UNIVERSIDAD AUTONOMA DE MADRID

ESCUELA POLITECNICA SUPERIOR

TRABAJO FIN DE MÁSTER

Identificación de Biomarcadores de Fibrilación Auricular Empleando Métodos Estadísticos e Inteligencia Artificial.

Máster Universitario en Bioinformática y Biología Computacional

Autor: Abdul Khalek Gharzeddine, Naim

Tutores:

Vera Rodríguez, Rubén
Departamento de Tecnología Electrónica y de las Comunicaciones

Ortega Rabbione, Guillermo José
Fundación para la Investigación Biomédica del Hospital Universitario de La Princesa.

FECHA: septiembre, 2020

ÍNDICE

ABREVIATURAS

- ECG: Electrocardiograma.
- FA: Fibrilación Auricular.
- RS: Ritmo Sinusal.
- IA: Inteligencia Artificial.
- SVM: Del inglés "*Support Vector Machine*".
- XGBoost: Del inglés "*eXtreme Gradient Boosting*".
- XML: Del inglés "*Extensible Markup Language*".
- ID: Número de identificación
- SFFS: Del inglés "*Sequential Forward Floating Selection*".
- pamp: Amplitud de la onda P.
- pdur: Duración de la onda P.
- parea: Área de la onda P.
- ppamp: Amplitud de la onda P prima.
- ppdur: Duración de la onda P prima.
- ppppdur: Duración de la onda P y P prima.
- pparea: Área de la onda P prima.
- pppparea: Área de la onda P y P prima.
- qamp: Amplitud de la onda Q.
- qdur: Duración de la onda Q.
- ramp: Amplitud de la onda R.
- rdur: Duración de la onda R.
- samp: Amplitud de la onda S.
- sdur: Duración de la onda S.
- rpamp: Amplitud de la onda R prima.
- rpdur: Duración de la onda R prima.
- spamp: Amplitud de la onda S prima.
- spdur: Duración de la onda S prima.
- vat: Tiempo de activación ventricular.
- qrsppk: Amplitud del complejo QRS de pico a pico.
- qrsdur: Duración del complejo QRS.
- qrsarea: Área del complejo QRS.
- ston: Elevación o depresión al inicio del Segmento ST.
- stmid: Elevación o depresión en el punto medio del segmento ST.
- st80: Elevación o depresión del segmento ST 80 milisegundos después del final del complejo QRS.
- stend: Elevación o depresión al final del segmento ST.
- stdur: Duración del segmento ST.
- stslope: Pendiente del segmento ST.

- tamp: Amplitud de la onda T.
- tdur: Duración de la onda T.
- tarea: Área de la onda T.
- tpamp: Amplitud de la onda T prima.
- tptpdur: Duración de la onda T y T prima.
- tpdur: Duración de la onda T prima.
- tparea: Área de la onda T prima.
- tptparea: Área de la onda T y T prima.
- print: Intervalo desde el inicio de la onda P hasta el inicio del complejo QRS.
- prseg: Intervalo desde el final de la onda P hasta el inicio del complejo QRS.
- qtint: Intervalo desde el inicio del complejo QRS hasta el final de la onda T.
- membercount: Número de latidos.
- memberpercent: Porcentaje del número total de latidos.
- longestrun: Corrida contigua más larga de latidos.
- meanqrsdur: Duración promedio del complejo QRS.
- lowventrate: Frecuencia ventricular más baja.
- meanventrate: Frecuencia ventricular promedio.
- highventrate: Frecuencia ventricular más alta.
- ventratestddev: Desviación estándar de la frecuencia ventricular.
- meanrrint: Intervalo promedio entre ondas R.
- atrialrate: Frecuencia auricular media.
- atrialratestddev: Desviación estándar de la frecuencia auricular.
- avgpcount: Número promedio de ondas P por complejo QRS.
- notavgpbeats: Número de complejos QRS que no tienen el número promedio de ondas P por complejo QRS.
- lowprint: Intervalo PR más corto.
- meanprint: Intervalo PR promedio.
- highprint: Intervalo PR más largo.
- printstddev: Desviación estándar del intervalo PR
- meanprseg: Segmento PR promedio.
- meanqtint: Intervalo QT promedio.
- meanqtseg: Segmento QT promedio.
- comppausecount: Número de latidos seguidos de una pausa compensatoria.

RESUMEN / ABSTRACT

RESUMEN

Las arritmias cardiacas tienen un peso considerable en la morbilidad y mortalidad en las enfermedades del corazón, generando más de un cuarto de millón de muertes al año en los Estados Unidos. Las arritmias pueden ocurrir durante la edad temprana, o pueden surgir más adelante debido a alguna enfermedad o el envejecimiento. La prueba más común utilizada para diagnosticar una arritmia es un electrocardiograma (ECG), el cual registra las diferencias de potencial eléctrico generadas por el corazón. Ciertas alteraciones en el patrón normal de la actividad eléctrica del corazón son indicativas de patologías cardíacas. Entre los distintos tipos de arritmias cardíacas la Fibrilación Auricular (FA) es la más común, y está asociada al envejecimiento. En el presente proyecto se analizaron más de 320.000 electrocardiogramas (ECGs) registrados en la base de datos del Hospital Universitario de La Princesa desde el año 2007 en formato XML, con el objeto de determinar biomarcadores y la generación de modelos predictivos de FA a partir de ECGs normales. Inicialmente se procedió con el estudio de la estructura de los archivos XML, y la identificación de la información de interés y sensible que pudiera identificar al paciente. Mediante un script en Bash la base de datos fue anonimizada, eliminando toda la información que pudiera identificar a los pacientes y generando nuevos números de identificación en una base de datos alterna. Posteriormente, con herramientas de análisis masivo se identificó, de forma anonimizada, aquellos pacientes que al menos tienen un ECG en FA y que a su vez presenten ECGs previos en Ritmo Sinusal (RS) normal (grupo de casos), al igual que pacientes que solo tienen registrados ECGs en RS (grupo control). El análisis masivo de más de 444 variables de ECGs en RS entre el grupo control y casos se llevó acabo por sexo y edad (de 40 a 49, 50 a 59, 60 a 69, 70 a 79, más de 80 años y el conjunto completo), y tomando en cuenta el tiempo entre ECGs. Una vez establecidos los grupos de estudio, se realizó un análisis estadístico para determinar si estos grupos presentaban diferencias significativas con respecto la edad, sexo y distancia entre ECGs, y se ajustaron para eliminar dichas diferencias. Seguido de esto, se llevó a cabo un análisis univariante para identificar de entre las más de 444 variables aquellas que presentan diferencias significativas entre casos y controles, y seguidamente con estas variables se construyeron modelos predictivos empleando los algoritmos de "*Extreme Gradient Boosting*" (XGBoost) y "*Support Vector Machines*" (SVM). Los resultados de exactitud obtenidos de estos ensayos se encuentran alrededor del 60%. Con el objeto de mejorar los resultados se empleó el método "*Sequential Forward Floating Selection*" (SFFS) o Selección secuencial flotante hacia adelante, el cual es otro método para la selección del conjunto de variables relevantes, obteniendo una mejoría en la exactitud del alrededor del 2%.

ABSTRACT

Cardiac arrhythmias have a considerable weight in morbidity and mortality in heart disease, generating more than a quarter of a million of deaths every year in the United States. Arrhythmias may occur at an early age, or they may arise later due to illness or aging. The most common test to diagnose an arrhythmia is an electrocardiogram (ECG), which records the differences in electrical potential generated by the heart. Certain alterations in the normal pattern of the electrical activity of the heart are indicative of cardiac pathologies. Among the different types of cardiac arrhythmias, Atrial Fibrillation (AF) is the most common, and is associated with aging. In this project, more than 320,000 electrocardiograms (ECG) registered in the database at the "Hospital Universitario de La Princesa" since 2007 in XML format were analyzed, in order to determine biomarkers and the generation of predictive models of AF from normal ECG. Initially, the structure of the XML files was studied, and the information of interest as well as all sensitive information that could identify the patient was detected. Through a script in Bash, the database was anonymized by removing all sensitive information that could identify the patient, new identification numbers were generated and an alternate database was created. Later, with massive data analysis tools, patients who have at least one ECG in AF and have a previous ECG in RS (cases group), as well as those patients who only have ECG in RS (control group) were identified and selected. The massive analysis of more than 444 variables of ECGs in SR between cases and control group were carried out by sex and age groups (from 40 to 49, 50 to 59, 60 to 69, 70 to 79, over 80 years old and the complete set), and taking into account the time between ECGs. Once the study groups were established, statistical studies were conducted to determine whether these groups had significant differences with respect to age, sex and distance between ECGs, and then were adjusted to eliminate such differences. Following this, a univariate analysis was performed to identify from all the variables in each group those which show significant differences between cases and controls, and then with these variables predictive models were constructed using the Extreme Gradient Boosting (XGBoost) and Support Vector Machines (SVM) algorithms. The accuracy of these models is around 60%. In order to improve the results, the Sequential Forward Floating Selection (SFFS) was used, which is another method for selection of a set of relevant variables, obtaining an overall improvement in the accuracy of 2%.

1. INTRODUCCIÓN

Las arritmias cardíacas son una causa importante de morbilidad y mortalidad en las enfermedades del corazón, y son la causa probable de más de un cuarto de millón de muertes anuales solo en los Estados Unidos. Las arritmias humanas se generan en dos posibles contextos: pueden ocurrir en la edad temprana, como consecuencia de problemas congénitos que ocasionan cambios en la anatomía y electrofisiología en el tiempo. Las arritmias también pueden surgir más tarde en la vida debido a alguna enfermedad adquirida o al envejecimiento (como la fibrilación auricular) [1].

Se pueden definir a las arritmias cardíacas como cualquier frecuencia o ritmo cardiaco anormal, y ocurren cuando no funcionan correctamente los impulsos eléctricos que coordinan los latidos del corazón, lo que trae como consecuencia que el corazón lata demasiado lento, demasiado rápido o de manera irregular. En adultos normales, el corazón late regularmente a un ritmo de 60 a 100 veces por minutos. Y el pulso coincide con las contracciones de los ventrículos [2].

1.1. MORFOLOGÍA Y FISIOLOGÍA DEL CORAZÓN.

El corazón está constituido por cuatro cámaras. Las dos cámaras superiores se llaman aurículas y las dos inferiores se denominan ventrículos. La aurícula izquierda y derecha están separadas por una pared muscular llamada tabique auricular, mientras que el ventrículo izquierdo y derecho están separados por el tabique ventricular (Figura 1). Durante el ciclo cardíaco las aurículas y los ventrículos se contraen y relajan produciendo un latido rítmico [3]. Cada latido se puede dividir en dos partes:

- Diástole: Es la fase del ciclo cardíaco que consiste en la relajación y llenado de las aurículas y ventrículos de sangre.
- Sístole: Es la fase del ciclo cardíaco en donde las aurículas se contraen (sístole auricular) y empujan la sangre hacia los ventrículos; luego, cuando las aurículas comienzan a relajarse, los ventrículos se contraen (sístole ventricular) y expulsan la sangre del corazón [3].

1.2. EL ELECTROCARDIOGRAMA.

La prueba más común utilizada para diagnosticar una arritmia es un electrocardiograma (ECG). El ECG son registros obtenidos de la superficie del cuerpo y registran las diferencias de potencial eléctrico generadas por el corazón. EL ECG usualmente es capaz de proporcionar un indicativo importante de una anomalía cardíaca e incluso permite hacer

una evaluación muy precisa de la anatomía y significado fisiológico de esa anormalidad, siendo el mejor método de análisis de las alteraciones del ritmo cardíaco [4].

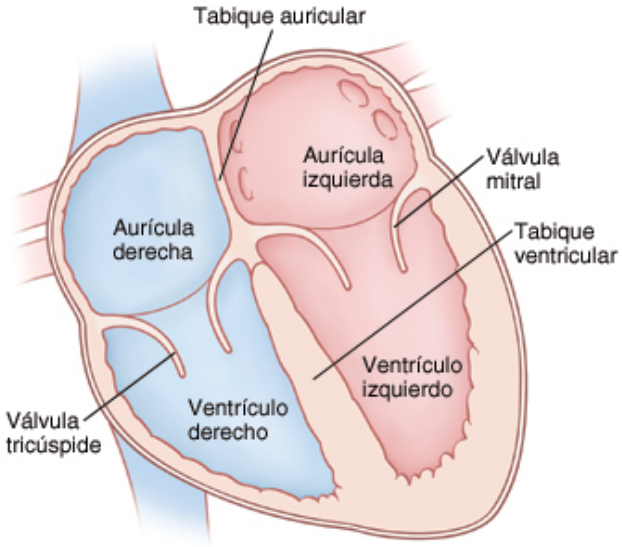

Figura 1. Estructura del corazón humano. Recuperado de http://healthlibrary.uchospitals.edu/

Un EGC estándar emplea 10 electrodos (6 en la pared torácica y 4 en las extremidades) para generar 12 vistas eléctricas del corazón conocidas como derivaciones. Estas derivaciones reflejan los ángulos en los que los electrodos "miran" el corazón y la dirección de la despolarización eléctrica del corazón [5].

Estas 12 derivaciones se dividen en 6 derivaciones provenientes de los miembros o extremidades y 6 derivaciones precordiales o torácicas. Las derivaciones de los miembros, también conocidas como derivaciones de plano frontal, generan grabaciones de derivaciones bipolares y unipolares aumentadas, mientras que las derivaciones torácicas o del plano horizontal o transversal registran derivaciones unipolares (Figura 2). Los registros de las derivaciones bipolares miden la diferencia de potencial entre dos electrodos, mientras que los registros unipolares, el electrodo de interés es comparado con uno de referencia [6].

Las derivaciones de las extremidades bipolares (Derivación I, II, III) miden las diferencias de potenciales en pares de electrodos en miembros:

- Derivación I (*Lead* I): Compara el brazo derecho (negativo) y el brazo izquierdo (positivo).
- Derivación II (*Lead* II): Compara el brazo derecho (negativo) con la pierna izquierda (positivo).

- Derivación III (*Lead* III): Compara el brazo izquierdo (negativo) con la pierna izquierda (positivo) [6].

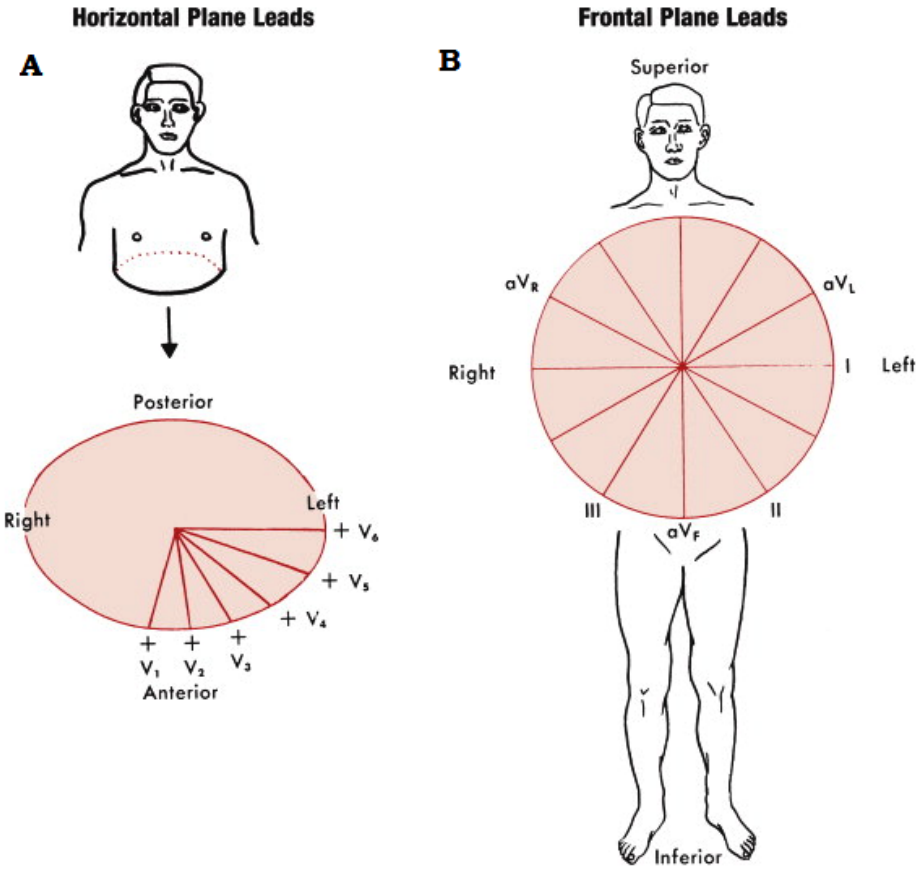

Figura 2. Derivaciones de plano frontal y horizontal (EGC estándar de 12 derivaciones). (**A**) Las relaciones espaciales de las seis derivaciones torácicas, que registran los voltajes eléctricos transmitidos al plano horizontal. (**B**) Las relaciones espaciales de las seis derivaciones de los miembros, que registran los voltajes eléctricos transmitidos al plano frontal del cuerpo. Recuperado de https://www.sciencedirect.com/topics/medicine-and-dentistry/ecg-leads

Las derivaciones de los miembros unipolares aumentadas (Derivación aVR, aVL y aVF) comparan el potencial de cada miembro con un electrodo de referencia:

- Derivación aVR (*Lead* aVR): El potencial del brazo derecho se compara con una referencia compuesta por los electrodos del brazo izquierdo y la pierna izquierda.
- Derivación aVL (*Lead* aVL): El potencial del brazo izquierdo se compara una referencia compuesta por los electrodos del brazo derecho y la pierna izquierda.
- Derivación aVF (*Lead* aVF): El potencial de la pierna izquierda se compara con una referencia compuesta por los electrodos del brazo derecho e izquierdo [6].

En la figura 3 se puede observar la configuración de los electrodos de las derivaciones de miembros o plano frontal.

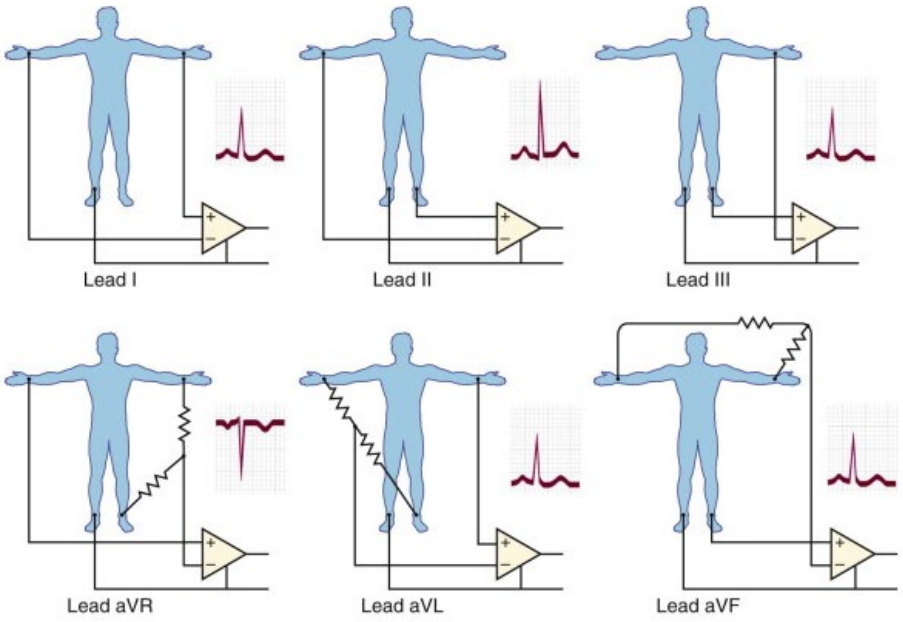

Figura 3. Configuración de los electrodos de las derivaciones de miembros. Recuperado de https://www.sciencedirect.com/topics/medicine-and-dentistry/ecg-leads

Los electrodos precordiales se colocan en puntos específicos de la pared torácica (Figura 4). Estas derivaciones comparan el potencial eléctrico entre el electrodo de tórax y un electrodo de referencia llamado "*Wilson central terminal"*, el cual combina los potenciales de brazo derecho, brazo izquierdo y pierna izquierda a través de resistencias de 5000 Ω [6].

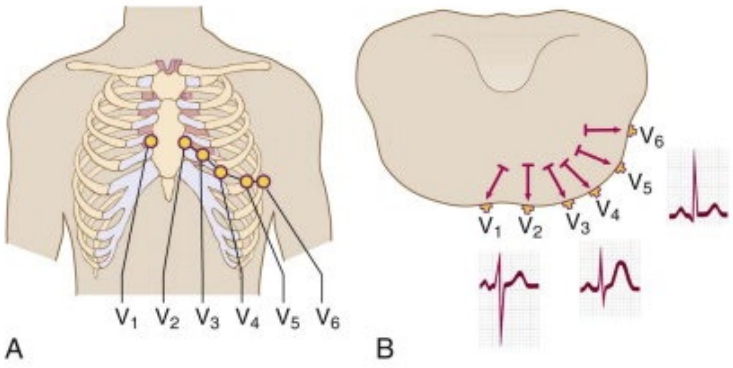

Figura 4. Derivaciones torácicas. (**A**) Posicionamiento de las derivaciones precordiales en la pared torácica. (**B**) Activación cardíaca normal como se manifiesta en las derivaciones precordiales. Recuperado de https://www.sciencedirect.com/topics/medicine-and-dentistry/ecg-leads

Las señales recogidas por los electrodos son procesadas para generar una representación gráfica de la actividad eléctrica del corazón. El patrón básico de esta actividad eléctrica, se compone de tres ondas, denominadas P, QRS (complejo de ondas) y T (Figura 5). La onda P es una pequeña onda que representa la despolarización auricular y el intervalo PR es el tiempo entre el principio de la onda P y el principio del complejo QRS. El complejo QRS representa la despolarización ventricular, las pequeñas ondas Q corresponden a la despolarización del tabique interventricular, la onda R representa la despolarización de la masa principal de los ventrículos y la onda S refleja la despolarización final de los ventrículos, en la base del corazón. El segmento ST, corresponde al segmento entre el final del complejo QRS y el inicio de la onda T, reflejando el período de potencial cero entre la despolarización ventricular y la repolarización. La onda T representa la repolarización ventricular. El intervalo QT, es el tiempo desde el principio del complejo QRS hasta el fin de la onda T. A veces se puede ver una onda U después de la onda T [5].

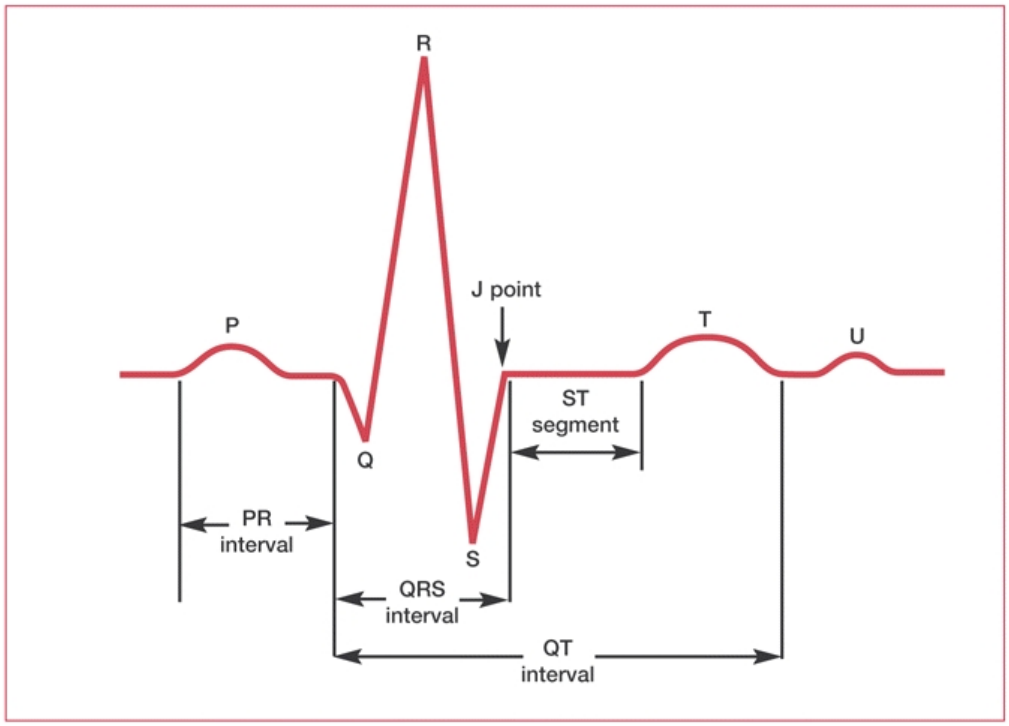

Figura 5. Patrón básico de la actividad eléctrica en el corazón. Recuperado de https://www.ncbi.nlm.nih.gov/books/NBK2214/

Es importante destacar que ciertas alteraciones en el patrón normal de la actividad eléctrica del corazón son indicativas de patologías cardíacas.

1.3. LA FIBRILACIÓN AURICULAR.

Entre las arritmias cardíacas la Fibrilación Auricular (FA) es la más común, y su frecuencia aumenta a medida que la población envejece [7]. En la FA se presenta una frecuencia

cardíaca irregular que ocurre cuando las dos cámaras superiores del corazón (aurículas) experimentan señales caóticas. Dentro de la cámara superior derecha del corazón se encuentran un conjunto de células llamadas nodo sinusal, las cuales son el marcapaso natural del corazón. El nodo sinusal genera una señal que normalmente inicia cada latido del corazón [8].

Esta señal que emite el nodo sinusal normalmente viaja a través de las dos cámaras superiores del corazón, y luego pasa por una vía de conexión entre las cámaras superiores e inferiores llamado nodo auriculoventricular (AV). El movimiento de esta señal hace que el corazón se contraiga, enviando sangre al corazón y el cuerpo. En la FA, las señales en las aurículas son caóticas, lo que da como resultado que estas tiemblen (Figura 6). Mientras que el nodo AV se encuentra bombardeado por impulsos que intentan llegar a los ventrículos, los cuales también laten rápidamente, pero no tan rápido como las aurículas, debido a que no pasan todos los impulsos [8].

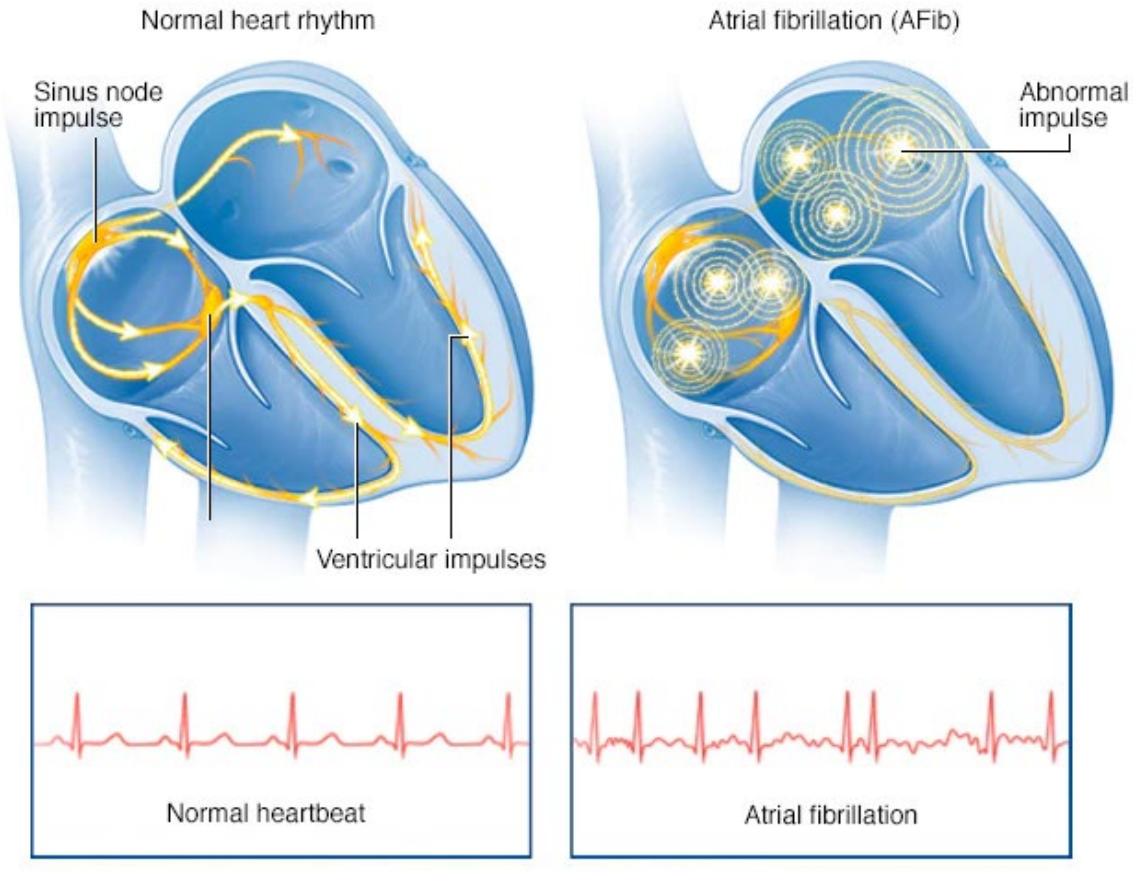

Figura 6. Comparación de un ritmo cardíaco normal y FA. Recuperado de https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624

La Fibrilación Auricular se puede clasificar en:

1. Paroxística: Es cuando los episodios terminan espontáneamente o con tratamiento dentro de los primeros 7 días. Pero pueden reaparecer con una frecuencia impredecible.
2. Persistente: Es cuando la FA es continua y dura más de 7 días, y no termina de forma espontánea.
3. Persistente de larga duración: Es cuando la FA continua dura más de 12 meses.
4. Permanente: Es cuando se acepta la FA y no se intentan tratamientos adicionales para restaurar o mantener el ritmo normal [9].

Posibles causas de la fibrilación auricular:

En ciertos casos la causa de la FA en personas es desconocida. Esta condición se conoce como FA solitaria o aislada. Usualmente, la fibrilación auricular es debida al daño del sistema de conducción eléctrico del corazón a causa de otras afecciones de salud, como:

- Bloqueo de una arteria pulmonar (embolia pulmonar).
- Enfermedad cardíaca congénita (defectos congénitos del corazón)
- Condiciones cardíacas, que incluyen un ataque cardíaco, insuficiencia cardíaca, miocardiopatía, enfermedad de las arterias coronarias o enfermedad cardíaca valvular.
- Cirugía del corazón.
- Inflamación de la membrana que rodea el corazón (pericarditis).
- Estrés por neumonía u otras infecciones.
- Problemas de tiroides, especialmente hipertiroidismo.
- Uso de algunos medicamentos, incluidos ciertos descongestionantes y píldoras de dieta.
- Uso de estimulantes como cafeína, tabaco, consumo excesivo de alcohol y algunas drogas ilegales [10].

Predisponentes a la fibrilación auricular:

A continuación se presentan los factores que pueden aumentar la posibilidad de padecer FA:

- Edad: El riesgo de desarrollar fibrilación auricular aumenta con la edad.
- Enfermedad cardiaca: Existe un mayor riesgo sufrir FA en personas con enfermedad cardíaca, tales como: problemas de las válvulas cardíacas, cardiopatía congénita, insuficiencia cardíaca congestiva, enfermedad de las arterias coronarias o antecedentes de ataque cardíaco o cirugía cardíaca.

- Hipertensión arterial: Tener presión arterial alta, especialmente si no está bien controlada con cambios en el estilo de vida o medicamentos, puede aumentar el riesgo de FA.
- Otras condiciones crónicas: Algunas afecciones crónicas como problemas de tiroides, apnea del sueño, síndrome metabólico, diabetes, enfermedad renal crónica o enfermedad pulmonar incrementan el riesgo de FA.
- Alcohol: Beber alcohol puede desencadenar un episodio de FA en algunas personas. Beber en exceso puede ponerlo en un riesgo aún mayor.
- Obesidad: Las personas obesas tienen un mayor riesgo de desarrollar FA.
- Historia familiar: Si se posee antecedentes familiares de FA hay un mayor riesgo de padecer esta condición [8].

Complicaciones de la Fibrilación Auricular:

- Derrame cerebral: Cuando las aurículas no bombean eficientemente existe el riesgo de formación de coágulos sanguíneos. Estos coágulos pueden moverse hacia los ventrículos y bombearse hacia el suministro de sangre, pasando a los pulmones o la circulación sanguínea general. Los coágulos en la circulación general pueden bloquear las arterias en el cerebro, causando un derrame cerebral. La FA aumenta el riesgo de accidente cerebrovascular entre 4 y 5 veces. Pero el riesgo depende de una serie de factores, incluida la edad y la presencia de hipertensión arterial, insuficiencia cardíaca, diabetes y antecedentes de coágulos sanguíneos.
- Insuficiencia cardiaca: En caso de FA persistente, el corazón puede comenzar a debilitarse. En casos extremos, puede provocar insuficiencia cardíaca, ya que el corazón no será capaz de bombear sangre alrededor del cuerpo eficientemente [11].

Siendo la FA una forma de arritmia cardíaca en la que el Ritmo Sinusal (RS) normal se reemplaza por despolarizaciones auriculares irregulares rápidas, esta es diagnosticada cuando hay ausencia de onda P y los complejos QRS aparecen a intervalos aleatorios irregulares [5].

En la Figura 7 se puede observar un ECG cuya morfología es indicativa de FA.

Figura 7. ECG que señala fibrilación auricular. Recuperado de https://www.ncbi.nlm.nih.gov/books/NBK2214/

1.4. ARCHIVOS XML.

El nombre XML proviene de sus siglas en inglés "*eXtensible Markup Language"*, traducido como "Lenguaje de Marcado Extensible" o "Lenguaje de Marcas Extensible". Un documento XML es un archivo de texto diseñado para almacenar datos de manera estructurada, como cualquier otro documento de texto, todo lo que hace es almacenar información. Podemos almacenar en formato XML documentos de Word, correos electrónicos, notas, etc. (Figura 8). Por ende, podemos considerar XML simplemente como un medio de almacenamiento destinado a ser un formato de intercambio de datos para sistemas que desean cooperar, proporcionando así las bases de la comunicación máquina a máquina [12].

Las partes más importantes de un documento XML son elementos y atributos. Los elementos consta de tres partes: una etiqueta de apertura, los datos en sí y una etiqueta de cierre. Usualmente se refiere a un elemento por su etiqueta de apertura. Es importante destacar que en XML no hay palabras reservadas, es decir, el nombre de un elemento puede ser una secuencia de caracteres arbitraria (observando algunas reglas simples, como que un nombre no puede comenzar con un dígito).

```
<?xml version="1.0" encoding="iso-8859-1"?>
<message>
   <from>Little Red Riding Hood</from>
   <to>Granny</to>
   <body>I'll visit you this afternoon!</body>
   <ps>I'll bring some cookies</ps>
   <ps>they say there's a wolf in the forest</ps>
</message>
```

Figura 8. Un mensaje de Little Red Riding Hood a Granny en formato XML. Fuente: [12]

En la Figura 8 se muestra un mensaje simple en formato XML. El elemento raíz del documento es <message>. Un documento XML debe contener exactamente un elemento raíz. Los elementos <from>, <to>, <body> y <ps> son todos hijos del elemento raíz. Dependiendo de su contenido, un elemento XML puede volverse complejo, mixto o simple. Los elementos complejos tienen hijos: como en el caso del elemento <message>. Un elemento simple contiene solo datos literales o numéricos; tal como <from> Little Red Riding Hood </from>. Un elemento mixto contiene tanto literales como elementos secundarios, mientras que un elemento vacío consiste solo en las etiquetas de apertura y cierre. La propiedad más llamativa de un documento XML es que almacena datos de manera jerárquica, definiendo un árbol, con el elemento raíz en su raíz. En la Figura 8 se detalla un árbol de dos niveles con una raíz y cinco elementos secundarios (Figura 9). Por

lo tanto, un documento XML establece no solo el nombre y el contenido de elementos individuales sino también sus relaciones jerárquicas. Esto hace que XML también sea adecuado para describir la estructura de los datos [12].

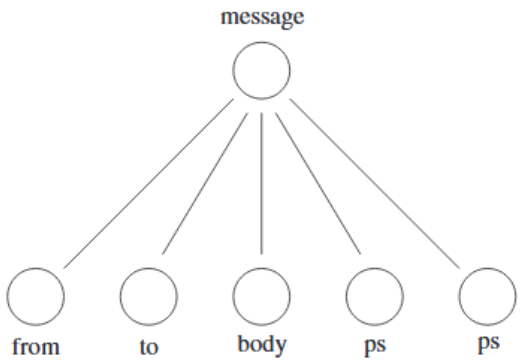

Figura 9. El árbol XML del mensaje de Little Red Riding Hood. Fuente: [12]

Los elementos XML pueden tener un número arbitrario de atributos (propiedades), pero cada atributo puede aparecer como máximo una vez en cada elemento. Un atributo tiene un nombre y un valor, separados por el signo de igualdad, con el valor entre comillas. En la Figura 10 se puede observar una representación XML alternativa del mensaje en la Figura 8. En donde el elemento <message> tiene como atributos al remitente y el destinatario del mensaje en lugar de ser elementos secundarios. Hay que tener en cuenta que un elemento con un atributo se considera complejo, pero los valores de los atributos solo pueden ser simples. El uso de atributos en lugar de elementos no tiene ningún beneficio sustancial. Esto es más una decisión de diseño y modelado. Los atributos pueden ser útiles en los casos en que solo contienen información auxiliar, no el núcleo de los datos. Su uso puede ayudar a que la fuente XML sea más legible y estructuralmente más limpia.

Cabe destacar, que los documentos XML pueden empezar con unas líneas que describen la versión XML, el tipo de documento, codificación, entre otras cosas. En la Figura 10 La primera línea específica que se trata de un documento XML y que utiliza la codificación de caracteres iso-8859-1 o Latin-1 [12].

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<message from="Little Red Riding Hood" to="Granny">
   <body>I'll visit you this afternoon!</body>
   <ps>I'll bring some cookies</ps>
   <ps>they say there's a wolf in the forest</ps>
</message>
```

Figura 10. Un mensaje de Little Red Riding Hood a Granny en formato XML y con atributos. Fuente: [12]

1.5. LA MEDICINA Y LA INTELIGENCIA ARTIFICIAL.

Actualmente, las organizaciones de atención médica enfrentan una presión creciente para lograr una mejor coordinación y mejorar los resultados de la atención al paciente. Para lograr esto, las organizaciones están recurriendo a la Inteligencia Artificial (IA) y el análisis predictivo [13]. La IA es el campo de estudio que intenta replicar las habilidades humanas, pero sin sus limitaciones de tiempo, energía y poder. Mediante el uso de algoritmos avanzados, las capacidades de procesamiento de datos y los sistemas de tecnología de la información se pueden producir predicciones basadas en datos en segundos, con poca o ninguna intervención humana.

El análisis predictivo utiliza tecnología, métodos estadísticos y algoritmos de aprendizaje automático o "*machine learning*" (parte de la IA) para buscar soluciones o respuestas a partir de cantidades masivas de información, analizándola y generando modelos predictivos, los cuales pueden ser aplicados a pacientes individuales. En medicina, las predicciones de los modelos de aprendizaje automático pueden variar desde respuestas a medicamentos hasta tasas de reingreso hospitalario. Los ejemplos incluyen predecir infecciones a partir de métodos de sutura, determinar la probabilidad de enfermedad, ayudar a un médico con un diagnóstico e incluso calcular el bienestar futuro [14]. Y por supuesto conociendo la morfología de la FA es posible crear modelos con el objeto de predecir esta patología.

Existen una gran variedad de algoritmos de aprendizaje automático, entre ellos se encuentran:

SVM ("*Support Vector Machine*")

Este es un algoritmo de aprendizaje automático supervisado que puede emplearse tanto para clasificación como para regresión. Los SVM para clasificación se basan en la idea de encontrar un hiperplano ("*Hyperplane*") que divida mejor un conjunto de datos en dos clases, siendo el hiperplano la línea que separa y clasifica linealmente estos datos. Los vectores de soporte ("*Support Vectors*") son los puntos más cercanos al hiperplano, los cuales si se eliminan, alterarían la posición del mismo. Debido a esto, pueden considerarse elementos críticos de un conjunto de datos. Intuitivamente, cuanto más lejos se encuentran el hiperplano de los datos, más seguro se está que se han clasificado correctamente. Por ende, es deseable que los puntos estén lo más lejos posible del hiperplano, sin dejar de estar en el lado correcto (Figura 11). La distancia entre el hiperplano y el punto más cercano de cualquiera de los conjuntos se conoce como el margen ("*Margin*"). El objetivo es elegir un hiperplano con el mayor margen posible, lo que brinda una mayor posibilidad de que los nuevos datos se clasifiquen correctamente [15].

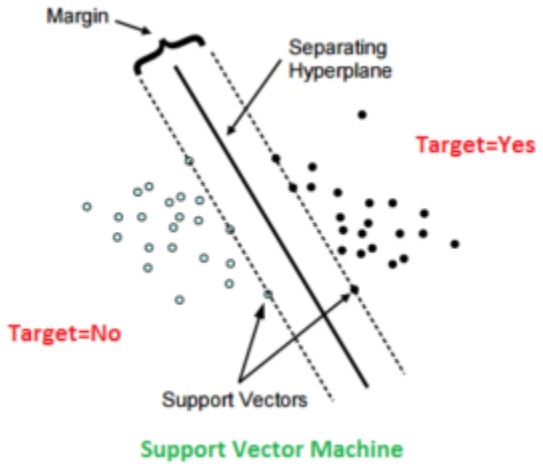

Figura 11. Ejemplo de clasificación con SVM y elementos del algoritmo. Recuperado de https://www.nosimpler.me/machine-learning/

Para el caso en el que el problema sea separable linealmente, SVM tratará de encontrar el hiperplano que maximice el margen, con la condición de que ambas clases se clasifiquen correctamente. Pero en realidad, los conjuntos de datos probablemente nunca sean linealmente separables, por lo que casi nunca se cumplirá la condición de que el 100% de los puntos sean clasificados correctamente por un hiperplano. Una de las formas en la que SVM aborda los casos que no son separables linealmente es mediante el "*kernel trick*". Lo que hace el "*kernel trick*" es utilizar las características existentes, aplicar algunas transformaciones y crear nuevas características, proyectando los datos en una dimensión más alta (Figura 12). La idea es que los datos continuarán siendo proyectados en dimensiones más altas y más altas hasta que se pueda formar un hiperplano para segregarlo [16].

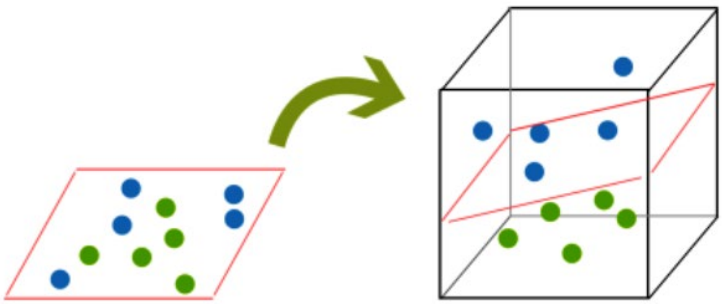

Figura 12. Ejemplo de clasificación con SVM empleando el "kernel trick". Recuperado de https://www.kdnuggets.com/2016/07/support-vector-machines-simple-explanation.html

XGBoost ("*eXtreme Gradient Boosting*")

XGBoost es un algoritmo de aprendizaje automático de ensamble ("*ensemble*") basado en árboles de decisión que usa "*gradient boosting*" como marco [17]. Los árboles de decisión son un método de aprendizaje supervisado no paramétrico utilizado para clasificación y regresión. Los árboles de decisión se basan en dividir un conjunto de datos en subconjuntos más pequeños basados en características descriptivas hasta que llegue a un conjunto lo suficientemente pequeño que contenga puntos que caigan bajo una misma etiqueta [18]. Cada característica del conjunto de datos se convierte en un nodo raíz o principal, y los nodos hoja o secundarios representan los resultados (Figura 13). La decisión sobre qué característica se usa para dividir se toma en base a la reducción de entropía resultante o la ganancia de información de la división [19].

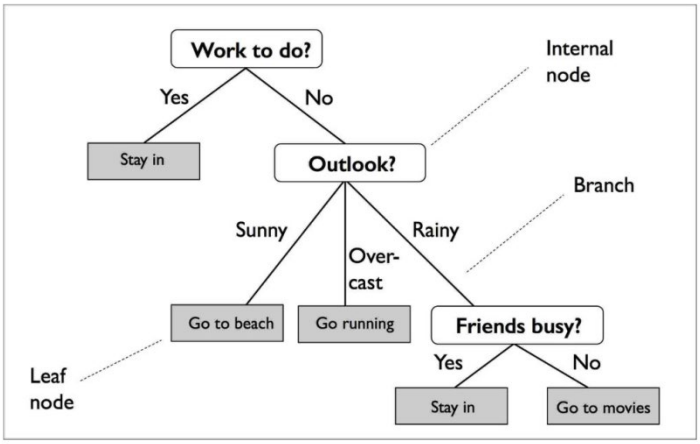

Figura 13. Ejemplo de árbol de decisión. Recuperado de https://towardsdatascience.com/https-medium-com-lorrli-classification-and-regression-analysis-with-decision-trees-c43cdbc58054

Por otro lado, "*Gradient Boosting*" es una técnica de aprendizaje automático para problemas de regresión y clasificación, que produce un modelo de predicción en forma de un ensamble de modelos de predicción débiles, típicamente árboles de decisión. Construye el modelo de manera escalonada como lo hacen otros métodos de refuerzo, y los generaliza al permitir la optimización de una función arbitraria de pérdida diferenciable. XGBoost es una de las implementaciones del concepto "*Gradient Boosting*", pero lo que hace que XGBoost sea único es que utiliza una formalización de modelo más regularizada para controlar el ajuste excesivo, lo que le da un mejor desempeño. Por lo tanto, ayuda a reducir el sobreajuste [20].

Tanto el algoritmo de SVM como XGBoost fueron empleados en el presente trabajo para analizar la información obtenida de una base de datos de más de 320.000 ECGs del Hospital Universitario de La Princesa localizado en Madrid. Esto con el objetivo de tratar de determinar marcadores y construir modelos que permitan predecir FA a partir de ECGs que son considerados normales.

2. OBJETIVOS

De forma general, el objetivo del presente trabajo es la identificación de marcadores predictivos y la construcción de modelos para el pronóstico de FA a partir del análisis de ECGs. Esto se llevó a cabo mediante el análisis masivo de una base de datos de más de 320.000 ECGs en formato XML almacenados desde el año 2007 en el Hospital Universitario de La Princesa. Y se establecieron los siguientes objetivos específicos:

- Evaluar la estructura de los archivos XML e identificar la información de interés.

- Anonimizar los archivos XML.

- Extraer la cuantificación del trazado así como también la señal de 10 segundos del ECG en cada una de sus 12 derivaciones.

- Filtrar la información para obtener 2 grupos poblacionales. El primero consta de pacientes los cuales en la base de datos tienen como mínimo un ECG en FA y un ECG previo normal. Y un segundo grupo de pacientes que tienen por lo menos 2 ECGs normales y que no tienen registrado ningún ECG en FA.

- Realizar el Análisis masivo de las más de 444 variables, empleando métodos estadísticos e inteligencia artificial, para la obtención de marcadores y modelos de predicción de FA.

3. METODOLOGÍA Y DESARROLLO

3.1. BASE DE DATOS.

En el presente proyecto se empleó una base de datos de más de 320.000 ECGs pertenecientes al Hospital Universitario de La Princesa localizado en el barrio de Salamanca de la ciudad de Madrid, España. Los archivos se encuentran en formato XML ("*Extensible Markup Language*") y fueron almacenados desde el año 2007 al 2019. En los archivos XML, se puede encontrar:

- La información personal del paciente.
- Los diferentes diagnósticos realizados para dicho ECG.
- La señal codificada y comprimida de cada una de las 12 derivaciones.
- Un conjunto de medidas cuantificadas a partir del análisis de las distintas señales, empleando para ello el algoritmo de 12 derivaciones de Philips.

El algoritmo de 12 derivaciones de Philips proporciona un análisis de las amplitudes, duraciones y morfologías de las formas de onda del ECG y el ritmo asociado. El análisis de la forma de onda del ECG se basa en criterios estándar para la interpretación de estos parámetros, los cálculos del eje eléctrico y la relación entre distintas derivaciones [21].

El algoritmo depende de la edad y el sexo, los cuales se utilizan para definir límites normales de frecuencia cardíaca, desviación del eje, intervalos de tiempo y valores de voltaje para la precisión de la interpretación del ECG. Los criterios para adultos se aplican si la edad del paciente es mayor o igual a 16 años, o si no se especifica la edad. Los criterios pediátricos se aplican si la edad del paciente es menor a 16 años.

Las declaraciones interpretativas del algoritmo de 12 derivaciones de Philips se producen mediante el uso de mediciones de ECG precisas y consistentes que son realizadas por el algoritmo. El proceso comienza con la adquisición simultánea de las doce derivaciones y posteriormente sigue cuatro pasos (Figura 14) para producir el informe del ECG interpretado:

1. Monitor de calidad: La calidad técnica de cada derivación del ECG es examinada. El análisis comienza obteniendo formas de onda de ECG precisas mediante la adquisición y el análisis simultáneo de la 12 derivaciones. Los datos de la forma de onda del ECG se capturan a una frecuencia de muestreo de 4 Mhz y se reducen a 500 muestras por segundo con una resolución de 5 μV. Durante el análisis, se estudia la calidad de la traza para garantizar buenas mediciones del ECG. El ECG también se analiza para detectar ruidos en la señal y poder corregirlos.

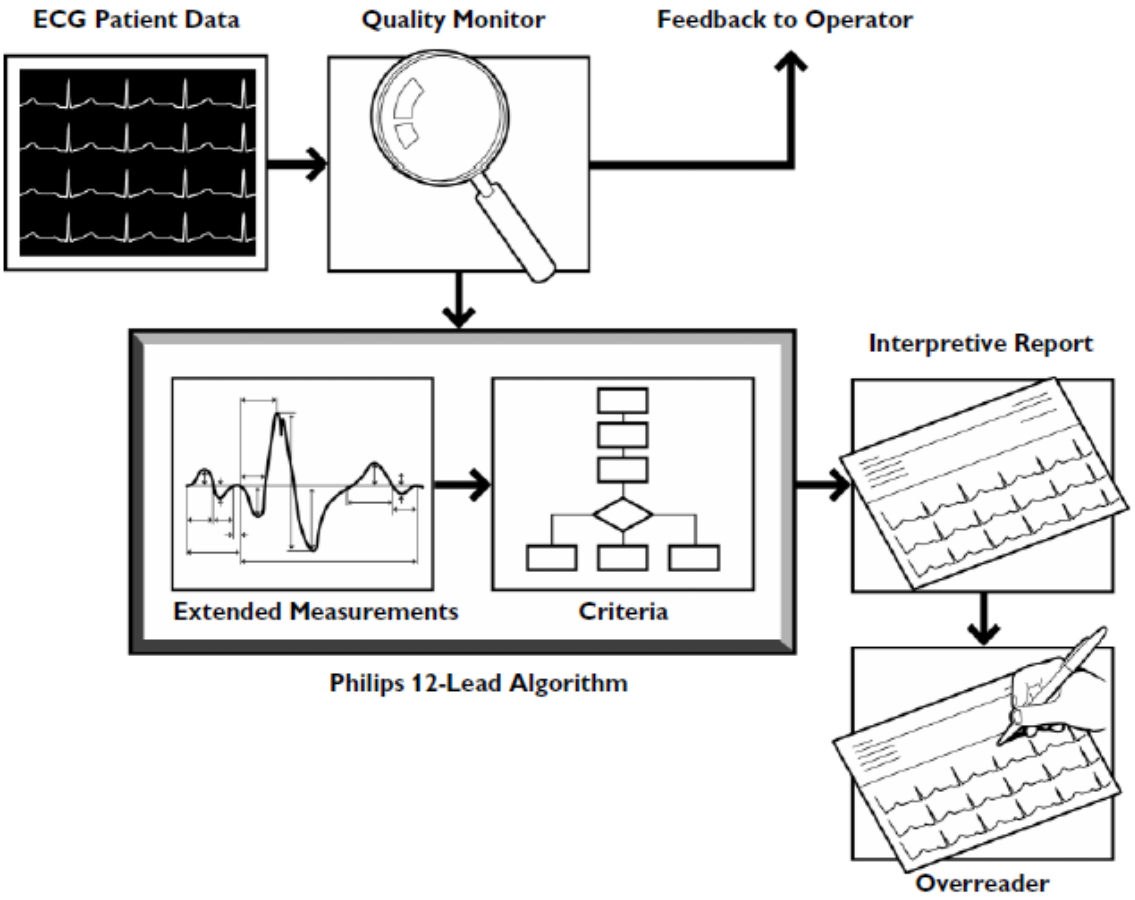

Figura 14. Pasos seguidos por el algoritmo de 12 derivaciones de Philips para producir el informe del ECG interpretado. Fuente: [21]

2. Reconocimiento de la forma de la onda: Localiza e identifica los diversos componentes de la forma de onda.

3. Medición: Mide cada componente de la forma de onda y realiza un análisis de ritmo básico, produciendo un conjunto completo de mediciones (Figura 15). Cada latido en cada derivación se mide individualmente, lo que permite que la variación natural entre los latidos contribuya a las mediciones representativas. En el algoritmo, todas las mediciones representativas de grupo, derivación y global se calculan a partir del conjunto integral de mediciones para cada latido. El algoritmo puede usar cualquier combinación de estos tres tipos de mediciones (grupal, de derivación, global), mejorando así la flexibilidad y el poder de sus capacidades interpretativas [21].

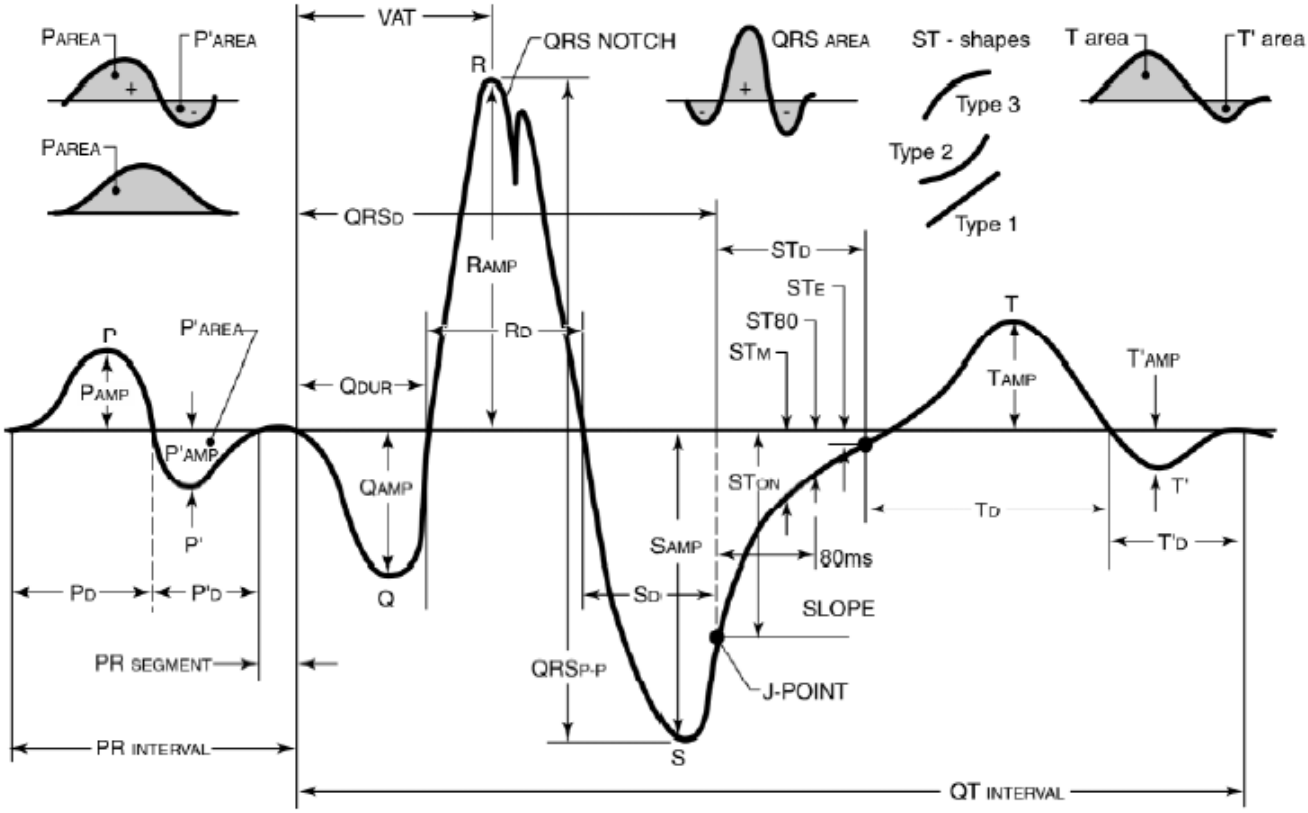

Figura 15. Mediciones de los componentes de la forma de onda del ECG. Fuente: [21]

Se realizan 3 tipos de mediciones:

- Mediciones grupales ("*Group Measurements*"): Cada latido en el ECG se clasifica en uno de los cinco grupos de ritmo según los parámetros de frecuencia y morfología. Cada grupo tiene ritmos con intervalos R-R, duraciones y formas similares. Las mediciones del grupo 1 representan el tipo de latido predominante. Las mediciones de los grupos 2 al 5 representan otros tipos de latidos cuyas medidas se promedian juntas.
- Mediciones de las derivaciones ("*Lead Measurements*"): Las mediciones para cada una de las 12 derivaciones se calculan a partir de los latidos del Grupo 1. Las mediciones de las derivaciones son promedios de la forma de onda dominante presente en cada derivación.
- Mediciones globales ("*Global Measurements*"): Estas mediciones de intervalo, duración y segmento son las mediciones del latido representativo en cada derivación del Grupo 1.

4. Interpretación: Utiliza diferentes mediciones e información del paciente (edad, sexo) para seleccionar declaraciones interpretativas [21].

3.2. ESTRUCTURA DE LOS ARCHIVOS XML E INFORMACIÓN DE INTERÉS.

En la Figura 16 podemos apreciar la estructura general de los archivos de los ECGs en formato XML del Hospital Universitario de La Princesa:

```
<?xml version="1.0" encoding="utf-16"?>
<restingecgdata xsi:schemaLocation="" status="" lang="" locale="" xmlns="" xmlns:xsi="">
  <documentinfo>
  <userdefines>
  <orderinfo priority="">
  <reportinfo date="" time="">
  <dataacquisition date="" time="" statflag="">
  <patient criteriaversionforpatientdata="">
  <internalmeasurements date="" time="" measurementversion="">
  <interpretations>
  <waveforms>
</restingecgdata>
```

Figura 16. Formato de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

En la primera línea se especifica la versión del archivo XML y la codificación que emplea, en este caso UTF-16. Como elemento raíz se encuentra <restingecgdata>, el cual está constituido por 9 elementos secundarios:

1. *"documentinfo"*
2. *"userdefines"*
3. *"orderinfo"*
4. *"reportinfo"*
5. *"dataacquisition"*
6. *"patient"*
7. *"internalmeasurements"*
8. *"interpretations"*
9. *"waveforms"*

Examinando el contenido de cada uno de estos 9 elementos secundarios se identificó en primer lugar, la información necesaria a extraer para llevar a cabo el estudio y en segundo lugar toda aquella información sensible que pudiera ser empleada para identificar al paciente.

En el elemento secundario "*documentinfo*", está compuesto a su vez por 6 elementos secundarios de los cuales 2 contienen información sensible que pudieran identificar al paciente: "*documentname*" y "*filename*". En estos elementos se encuentra el nombre del archivo y su ubicación (Figura 17).

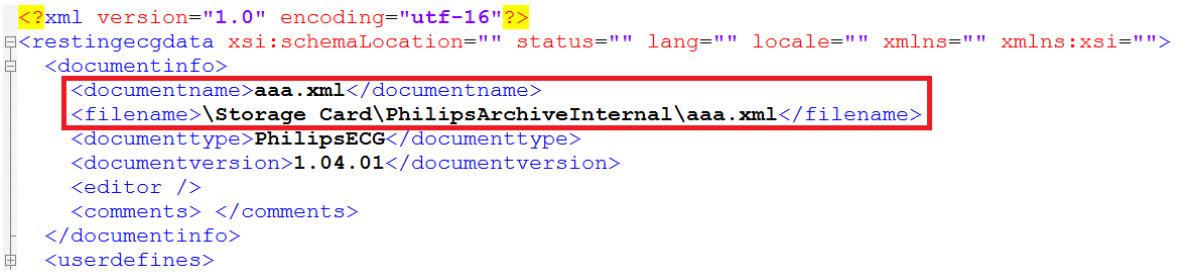

Figura 17. Estructura del elemento secundario "*documentinfo*", de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

En los elementos secundarios *"userdefines"*, *"orderinfo"* y *"reportinfo"* no se encontró ninguna información de interés para este estudio en particular. En cuanto al elemento *"dataacquisition"*, está constituido por los elementos *"machine"*, *"acquirer"* y *"signalcharacteristics"*. En los atributos de *"dataacquisition"* se puede encontrar información relevante como lo es la fecha *("date")* y la hora (*"time")* en la que se realizó el estudio en el paciente. Mientras que en el elemento *"machine"* se encuentran los atributos *"machineid"* y *"detaildescription"* los cuales contienen información sensible a ser removida (Figura 18).

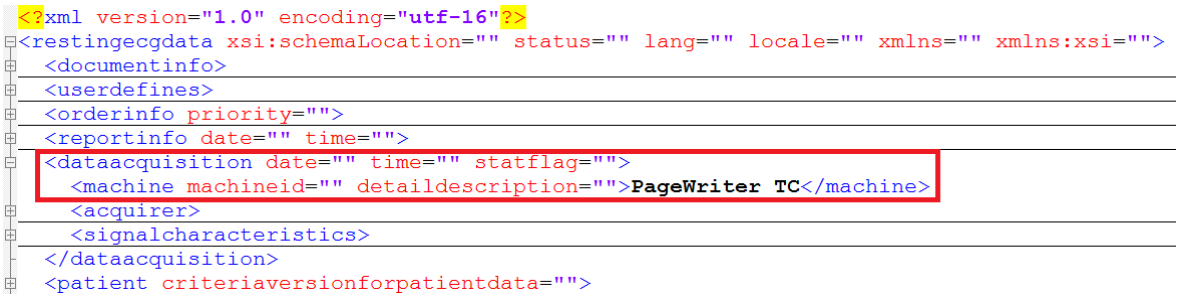

Figura 18. Estructura del elemento secundario "*documentinfo*", de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

Con respecto a elemento secundario *"patient"* (Figura 19) está constituido por dos elementos: *"generalpatientdata"* y *"patientmedicaldata"*. En este elemento está contenida la información del paciente, por lo tanto habrá información necesaria para el estudio e información sensible que no deberá estar presente en los archivos anonimizados. Es importante destacar que la información contenida en este elemento es introducida manualmente, por lo tanto es propensa a presentar errores, o también es posible que existan valores ausentes.

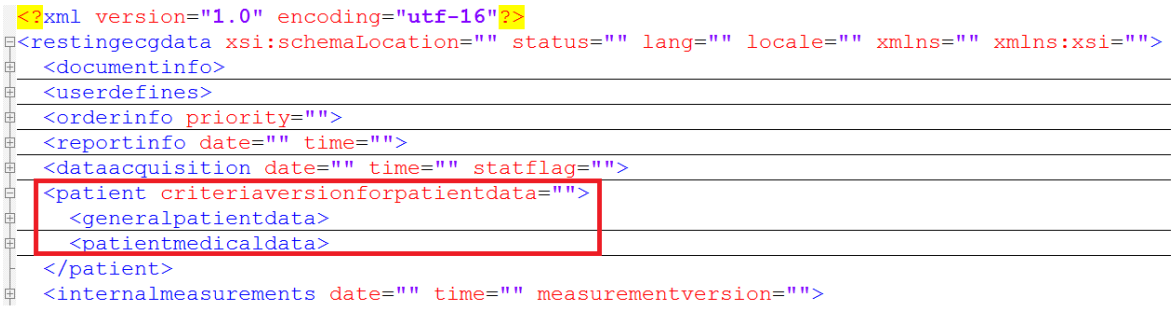

Figura 19. Estructura del elemento secundario "*patient*", de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

En la figura 20 podemos apreciar la estructura del elemento secundario *"generalpatientdata"* perteneciente al elemento *"patient"*, el cual a su vez está formado por 9 sub-elementos: *"patientid"*, *"uniquepatientid"*, *"MRN"*, *"name"*, *"age"*, *"pacestatus"*, *"sex"*, *"height"*, *"weight"*. Entre estos elementos los siguiente son los que suelen contener información y son de interés:

1. *"patientid"*: ID único o número de identificación del paciente.
2. *"name"* (*"lastname"* y *"frstname"*): Apellido y nombre del paciente.
3. *"age"*: Edad del paciente.
4. *"sex"*: Sexo del paciente.

Por otra parte, en la figura 21 podemos apreciar la estructura del elemento secundario *"patientmedicaldata"* perteneciente al elemento *"patient"*, constituida por un solo elemento secundario "*bloodpressure*", normalmente la información es esta sección no está complementada, por lo tanto no se tomará en cuenta para el estudio.

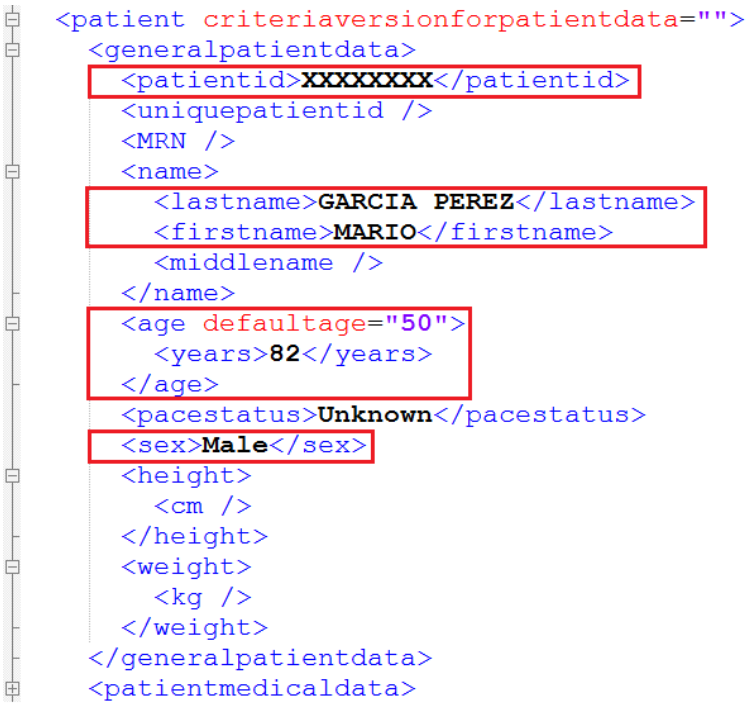

Figura 20. Estructura del elemento secundario *"generalpatientdata"* perteneciente a "*patient*", de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

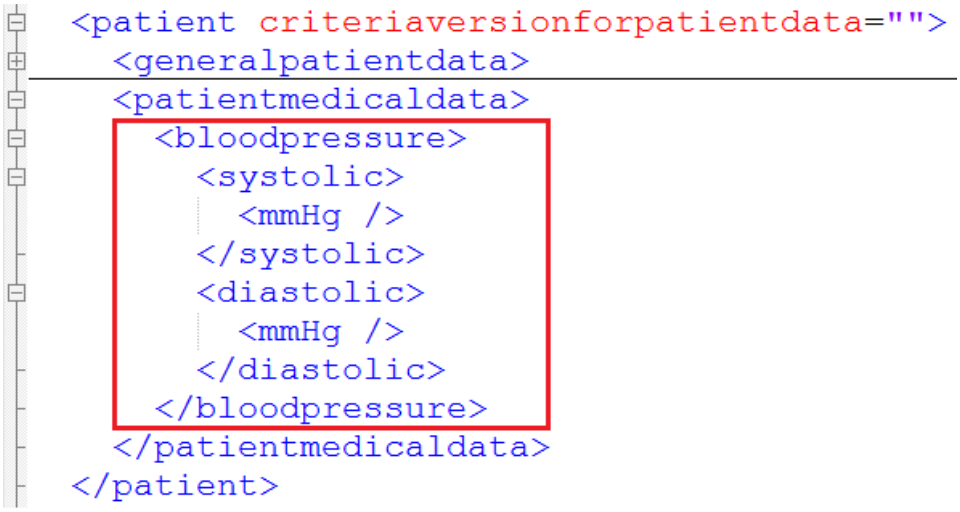

Figura 21. Estructura del elemento secundario *"patientmedicaldata"* perteneciente a "*patient*", de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

En el elemento *"internalmeasurements"*, se pueden encontrar los valores de las variables cuantificadas por el algoritmo de 12 derivaciones de Philips a partir de las ondas de las distintas derivaciones.

Es importante destacar que el elemento *"internalmeasurements"*, puede no estar presente en el archivo XML, debido a que para que el algoritmo de Philips realice las respectivas cuantificaciones es necesario que así se indique, es decir que no se realizan automáticamente.

Para el elemento *"interpretations"* (Figura 22), se pueden encontrar los siguientes elementos secundarios:

1. *"interpretationdatastructure"*: En esta sección se hayan las interpretaciones hechas por el algoritmo de Philips en base a las mediciones realizadas.
2. *"globalmeasurements"*: Mediciones globales del ECG.
3. *"mdsignatureline"*: Confirmación de la interpretación hecha por el algoritmo por parte del médico.
4. *"severity"*: Diagnóstico de la severidad general hecha por el algoritmo.
5. *"statement"*: Presentación de las diferentes interpretaciones hechas por el algoritmo de forma individual, estas mismas interpretaciones se pueden localizar en el elemento *"interpretationdatastructure"*. Es importante destacar que el número de interpretaciones hechas por el algoritmo varía en los archivos XML, por lo tanto también lo hará el número de elementos *"statement"* en cada archivo XML.

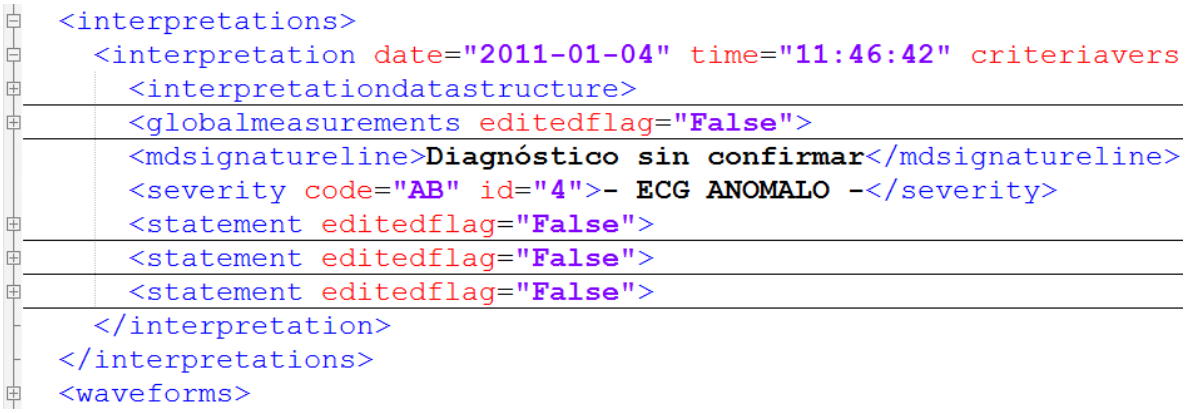

Figura 22. Estructura del elemento secundario *"internalmeasurements"* de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

Finalmente, en el elemento *"waveforms"* se encuentran las ondas de las 12 derivaciones codificadas y comprimidas.

3.3. ANONIMIZACIÓN DE LOS ARCHIVOS XML.

Una vez identificada la información sensible y su ubicación, se procedió al desarrollo de un script en Bash para remover dicha información en cada uno de los archivos XML. Este script requiere de tres argumentos:

1. La dirección de la carpeta donde se almacenaran los archivos anonimizados.
2. La dirección de la carpeta donde se encuentran los archivos a anonimizar.
3. La dirección del documento en el cual se mantiene el registro de los datos de los pacientes. Este documento permite la identificación de los archivos anonimizados, en caso de que así fuese necesario, como por ejemplo para indagar con mayor profundidad sobre pacientes específicos que sean de particular interés. Cabe destacar que a este documento solo tendrán acceso el personal autorizado.

El programa creado sigue la siguiente secuencia de órdenes (Figura 23):

1. Verifica si existe la carpeta donde se almacenaran los archivos anonimizados. En caso de que esta carpeta no exista la crea.

2. Verifica que el documento donde se registran los datos exista, en caso de que no exista lo crea, para referenciar este documento se le dará el nombre de "datos.txt". Este documento consta de 6 columnas, las cuales están separadas por tabulaciones:

    - Columna 1: Nombre del paciente.
    - Columna 2: Apellido del paciente.
    - Columna 3: ID o número de identificación del paciente.
    - Columna 4: Nuevo ID del paciente.
    - Columna 5: Nombre del archivo XML original.
    - Columna 6: Nuevo nombre del archivo XML.

3. Verifica la existencia de un archivo llamado "id_errors.txt", en caso de que no exista lo crea. Ya que los ID son introducidos manualmente es posible que existan errores en los mismos, por lo cual se genera un archivo automático para mantener un registro de estos, y que posteriormente se puedan corregir en la bases de datos del hospital. Este archivo también consta de 6 columnas separadas por tabulaciones:

    - Columna 1: El número del ID con error.
    - Columna 2: Error presente en el ID (error en el formato o ausente)
    - Columna 3: Nombre del paciente.
    - Columna 4: Apellido del paciente.
    - Columna 5: Nombre del archivo XML original.

- Columna 6: Nuevo nombre del archivo XML.

4. A partir de este punto se inicia un bucle donde se procesan cada uno de los archivos XML contenidos en la carpeta de archivos a anonimizar.

5. Se crean dos archivos temporales: uno contiene los ID originales de "datos.txt" y se llama "tmp_ID.txt", el otro archivo contiene los nuevos ID generados y se llama "tmp_new_ID.txt". Estos archivos se sobrescriben en cada iteración de forma que se actualizan constantemente.

6. Se verifica si el archivo es UTF-16, de ser el caso se transforma a UTF-8 para poder ser procesado en la terminal de Linux, y se almacena con su nombre original en la carpeta de archivos anonimizados.

7. Con el comando awk se extraen las líneas donde se encuentran la información de interés y el número de cada línea, las cuales se almacena en un archivo temporal, llamado "tmp_inf.txt", para posteriormente ser procesado. Para buscar la información se emplean los nombres de las etiquetas correspondientes anteriormente identificadas, las cuales son únicas:

   - *"patientid"*
   - *"dataacquisition"*
   - *"firstname"*
   - *"lastname"*
   - *"machine"*
   - *"documentname"*
   - *"filename"*

8. En caso de que el ID se una cadena numérica ininterrumpida, este se almacena en una variable, al igual que el número de la línea donde se encuentra en el archivo. Para identificar y separar la información puntual se usa el comando sed con expresiones regulares, extrayendo solamente el ID de la línea, al igual que el número de línea. Si el ID no es una cadena numérica ininterrumpida o el campo está vacío, este archivo no se sigue procesando, se registra en el documento "id_errors.txt", se remueve el XML de la carpeta de anonimizados, y se salta al siguiente archivo.

9. De las líneas donde se encuentran las etiquetas *"firstname", "lastname", "machine", "documentname", "filename"* en el archivo "tmp_inf.txt", se extrae la información de interés y el número de la línea donde se encuentra esta información

en el archivo XML, los cuales se almacenan en variables. Luego se elimina dicha información del archivo especificando su ubicación exacta, no solo la línea sino también su localización dentro de la línea en particular, esto para evitar realizar modificaciones que pudiera alterar la información y estructura del archivo XML de manera indeseada. Para ello se emplearon los comandos sed y awk además de expresiones regulares y grupos dentro de dichas expresiones.

10. A partir de la línea con la etiqueta *"dataacquisition"* en el archivo "tmp_inf.txt", se extrajo y almaceno en variables la fecha y hora en la que se realizó el ECG, empleando para ello el comando awk.

11. Se verifica si el ID previamente guardado se encuentra registrado en "tmp_ID.txt" (registro de IDs). En caso de que ya este registrado, a partir del ID original se busca el nuevo ID en "datos.txt". Se asigna un nuevo nombre al archivo XML el cual está constituido por: el nuevo ID, la fecha y la hora. Se modifica el nombre del archivo en la carpeta de archivos anonimizados, se modifica el ID en el documento, y se registra el archivo en "datos.txt".

12. Si el ID no se encuentra registrado, se obtiene un nuevo ID empleando para ello /dev/urandom que es un archivo especial que sirve como generador de números seudo-aleatorios. Una vez creado, verifica si existe en "tmp_new_ID.txt" (registro de nuevos IDs), si existe genera otro número de identificación y verifica nuevamente que no exista, así hasta obtener un nuevo número. Luego se genera el nuevo nombre del archivo, constituido por: el nuevo ID, la fecha y la hora. Se modifica el nombre en la carpeta de anonimizados, se cambia el ID original por el nuevo ID en el archivo XML, y se registra el archivo en "datos.txt".

13. El proceso se repite hasta que se procesan todos los archivos en la carpeta de archivos a anonimizar.

14. Finalmente se eliminan los archivos temporales.

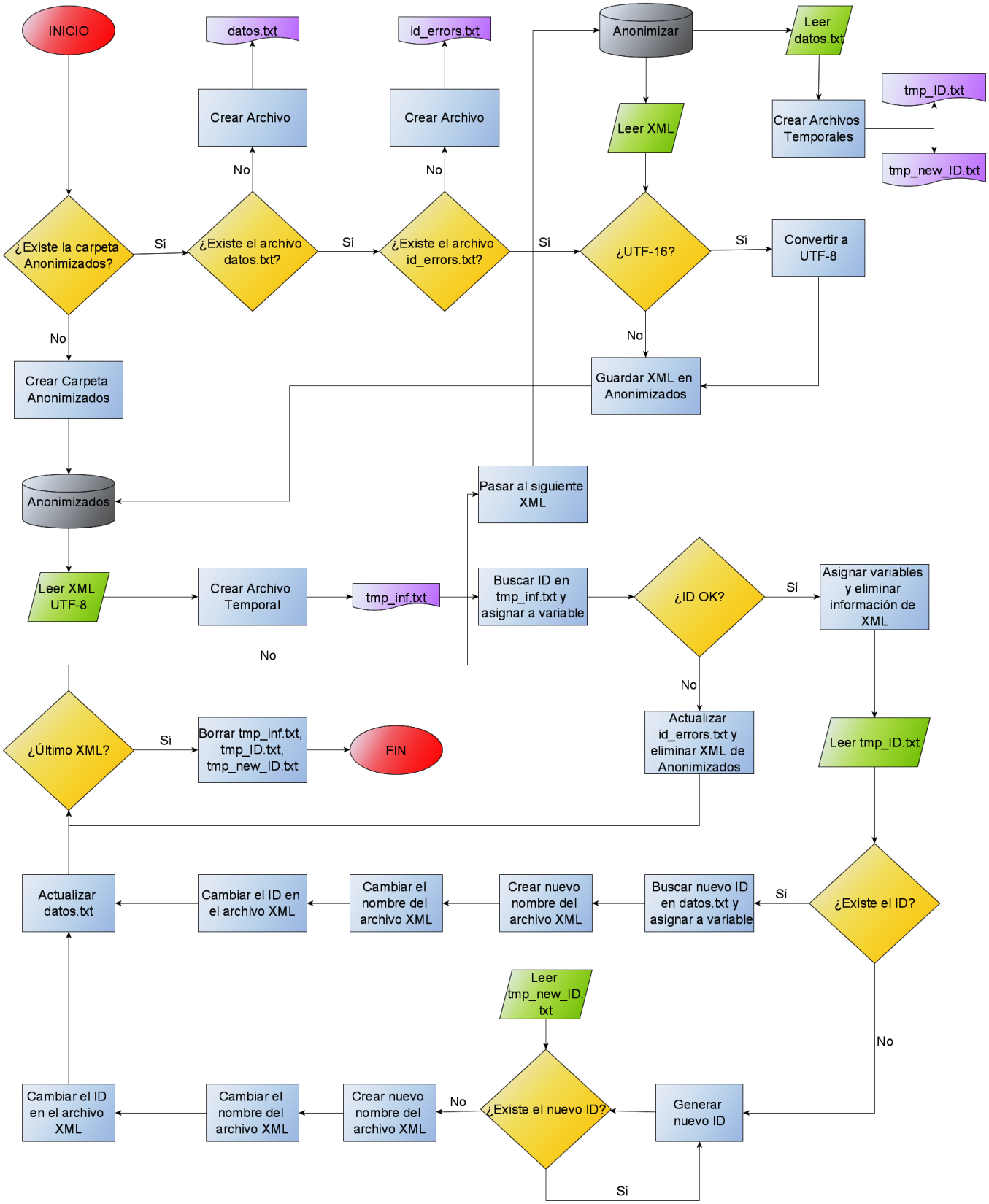

Figura 23. Diagrama de Flujo del script para la anonimización de los archivos XML.

3.4. EXTRACCIÓN DE LAS CUANTIFICACIONES DEL TRAZADO HECHAS POR EL ALGORITMO DE PHILIPS.

Para extracción de la información de los archivos XML anonimizados se desarrolló un script en R, empleando el paquete XML [22], este paquete consiste en un conjunto de herramientas para extraer y analizar la información contenida en los XML dentro de R y S-Plus. De este paquete se emplearon las siguientes funciones:

- xmlParse: Analiza un archivo o cadena XML/HTML y genera una estructura en R que representa el árbol XML / HTML.

- xmlRoot: Se trata de una colección de métodos para proporcionar un acceso fácil al objeto XMLNode de nivel superior resultante del análisis de un documento XML (xmlParse).

- xmlSize: Esta función determina el número de elementos secundarios dentro de un elemento o nodo específico.

- xmlValue: Proporciona acceso al contenido bruto de un elemento o nodo.

- xmlName: Devuelve el nombre de la etiqueta de un elemento o nodo.

- xmlAttrs: Devuelve un vector que proporciona los atributos de un elemento o nodo.

En el script se llevan a cabo los siguientes pasos:

1. Inicialmente se procedió a leer los archivos XML presentes en la carpeta y almacenar los nombres de los mismos en una variable.

2. Se crean diferentes listas y vectores para almacenar la información.

3. Se abre un bucle para procesar todos los archivos XML. A partir de este paso las acciones se ejecutaran para cada uno de los archivos.

4. El archivo se lee con xmlParse para crear el árbol del XML en R, y luego con la función xlmRoot se determina el elemento raíz que posteriormente permitirá acceder a los demás elementos del árbol.

5. Se identifican el número de los nodos donde se encuentra la información a extraer:

    - Fecha y hora: se encuentra en el nodo 5 (*"dataacquisition"*).

- Datos del paciente: nodo 6 (“*patient”*).
- Datos cuantificados de las ondas: nodo 7 (“*internalmeasurements”*).
- Diagnóstico: nodo 8 (“*interpretations”*).
- Señal de la onda: nodo 9 (“*waveforms”*).

6. Para extraer los datos de fecha y hora se accede al nodo 5, y con la función xmlAttrs se obtiene la información del mismo.

7. Para el resto de la información se accede al nodo específico donde se encuentra. Primeramente se determina el tamaño de cada nodo en particular (con xmlSize) y a partir de su tamaño se ejecuta un bucle para recorrer cada uno de sus elementos extrayendo el valor del elemento y su nombre, con las funciones xmlValue y xmlName respectivamente. Esto valores son almacenados en listas y vectores.

8. Para el caso particular del nodo “*internalmeasurements”*, es posible que se encuentre ausente en el archivo XML, ya que como se mencionó anteriormente no se genera de forma automática. En este caso se verifica que esté presente, de no ser así, se asigna el valor 0 a las variables de este archivo en las listas correspondientes. Además se registra el índice de dicho archivo en un vector para luego poder identificar al mismo y excluirlo de la base de datos, pues no tiene ninguna utilidad para el estudio.

9. Igualmente para el nodo “*internalmeasurements”* es necesario hacer otro bucle interno para extraer los valores correspondiente a cada una de las 12 derivaciones (I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6) presente en los archivos XML, generando un array de 3 dimensiones.

10. La información extraída es organizada en dataframes de 2 y 3 dimensiones, donde las filas representan cada paciente y las columnas las diferentes variables, y para el caso de “*internalmeasurements”* la tercera dimensión representa cada derivación.

11. Con respecto a la información contenida en *“interpretations”* tiene un número variable de sub-elementos *“statement”*, por lo cual es necesario determinar el número máximo de este elemento, y emplear este valor en la configuración del dataframe. Para el caso de los archivos que tengan un número inferior del elemento *“statement”* se asigna el valor 0 en el campo correspondiente a las columnas ausentes.

3.5. SELECCIÓN DE LOS PACIENTES Y ARCHIVOS XML.

Para la selección de los registros se emplea la información de los pacientes anonimizados y el diagnóstico correspondiente a cada archivo XML. El objetivo de este paso es obtener dos grupos de pacientes: casos y controles. El grupo de casos está constituido por aquellos individuos que tenga registrados en la base de datos (diagnostico automático del algoritmo) por lo menos un ECG en FA y otro ECG previo en RS. El grupo control está constituido por aquellos que presentan como mínimo 2 ECG normales y que no tengan registrado ningún ECG en FA. Para este proceso se siguieron los siguientes pasos:

1. Con la totalidad de los datos originales se determinó la prevalencia de FA y se comparó con la bibliografía (Figura 24) :

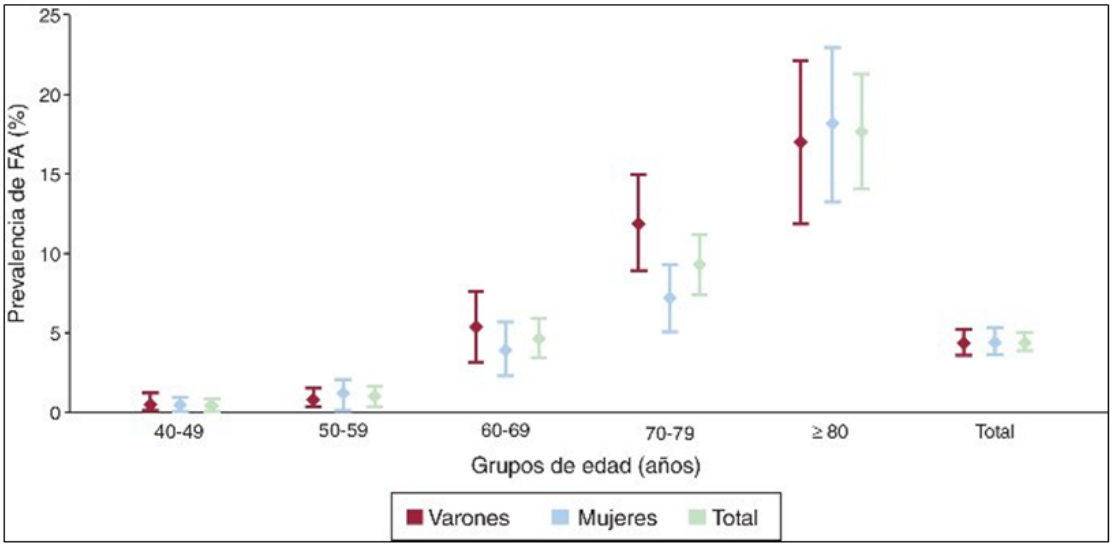

Figura 24. Representación gráfica de la prevalencia de FA en España resultante del estudio OFRECE. Fuente: [23].

La prevalencia se entiende como la proporción de una población que tiene una característica específica en un período de tiempo determinado [24]. Como se puede observar en la Tabla 1 los resultados de prevalencia para los ECGs de la base de datos del Hospital Universitario de La Princesa son similares a los obtenidos en estudios previos [23] representados en la Figura 24.

Tabla 1. Resultados de prevalencia de FA obtenidos de los ECGs de la base de datos del Hospital Universitario de La Princesa.

	Prevalencia de FA (%)				
	40-49	**50-59**	**60-69**	**70-79**	**>80**
Hombres	2.13	4.15	8.74	17.17	28.42
Mujeres	1.44	2.02	6.02	12.63	25.66
Hombres y Mujeres	1.79	3.03	7.42	14.76	26.73

2. Se eliminaron todos aquellos registros que no tenían las cuantificaciones del trazado de las ondas. Aquí se produjo una reducción en el número de registros de 327.069 a 301.277, y con una disminución en el número de pacientes de 123.811 a 117.536.

3. El siguiente paso en el filtrado fue la eliminación de todas las personas que solo presentan un ECG en la base de datos, ya que para el diseño experimental de este estudio no son útiles. En este paso se produjo una reducción de 66.221 ECG correspondiente al mismo número de pacientes, resultando en un total de 235.056 ECG perteneciente a 51.315 pacientes.

4. Se crearon dos dataframes: uno compuestos por todos los ECGs en FA y otro con todos los ECGs en RS. Los dataframe en FA están constituidos por 34.769 registros de 10.496 pacientes, mientras que el dataframe en RS presenta 157.235 registros de 45.634 pacientes. Teniendo una población total de 49.370 pacientes (hay pacientes presentes en ambos dataframes) y 192.004 ECGs.

5. Un mismo paciente puede tener más de un ECG en FA, por lo cual se seleccionó el ECG más antiguo. Arrojando un total de 10.496 ECGs para el mismo número de pacientes. El número total de ECGs se reduce a 167.731 y el número de pacientes se mantiene igual, 49.370.

6. Agrupados los registros con diagnostico FA y SR, se procedió a identificar aquellos pacientes que habían sido diagnosticados con FA y tienen por lo menos un ECG previo en RS (Grupo FA). Para ello se realizó un bucle que itera por todos los ID del dataframe FA, cada uno de esos ID se emplea para buscar los registros en el dataframe RS. En caso de existir, se verifica que la fecha del registro FA sea posterior al registro RS, y de ser así, tanto el registro FA como el o los registros RS se almacenan en un nuevo dataframe. Este nuevo dataframe incluye 2 nuevas variables: el número de registros por paciente, y la distancia en días del registro FA con respecto al/los registro(s) RS. El Grupo FA está constituido por 4.466 pacientes con 4.466 ECGs en FA y 13.301 ECGs en RS (17.767 ECGs en total)

7. Igualmente, se identificaron aquellos pacientes que en la base de datos tienen registrados dos a más ECGs en RS y no tienen ninguno en FA (Grupo RS). Nuevamente, se realizó un bucle con los ID únicos en el dataframe RS, se verifico que dichos ID no aparecen en el dataframe de FA. De ser el caso, se verifica que hayan por los menos dos registros, si es así se almacenan en un nuevo dataframe. Si solo existe un registro en RS o si el paciente también tiene un registro en FA tanto el paciente como los ECG son descartados. El grupo RS está constituido por 32.238 pacientes con 118.595 ECGs en RS. A partir de estos 2 últimos procesos de filtrado el número de ECGs se redujo a 136.362 y el número de pacientes a 36.704.

8. Dado que la mayoría de los pacientes presenta 2 ECGs, se seleccionó el mismo número de ECGs para cada paciente, asegurando que la información fuera igualmente representativa y el estudio consistente. La selección se realizó tomando en cuenta la distancia media entre EGCs, la cual se calculó para cada dataframe (FA y RS). Se tomó el último ECGs de cada individuo, y luego se seleccionó el ECG que tuviera la menor distancia con la media. En este paso se mantuvo el mismo número de pacientes, y con respecto a los registros, para el Grupo FA hay 4.466 ECGs en FA y 4.466 ECGs en RS (8932 ECGs en total) y para el Grupo RS hay 64.476 ECGs en RS

9. Dado que la información del paciente se introduce de forma manual es posible que hayan datos que presenten errores o estén ausentes. La calidad del ID se verifica durante el proceso de anonimizado, el nombre y apellido no se emplean en esta parte del proceso, por lo que restaría la verificación de la edad y el sexo. Tras este último filtrado la población final sería 29.353 (58.706 ECGs), donde 3.334 pacientes están en el grupo FA con 6.668 ECGs (3.334 en FA y 3.334 en RS) y 26.019 pacientes están en el grupo RS con 52.038 ECGs en RS.

En la Figura 25 se puede observar el diagrama de la selección de pacientes y archivos XML a ser analizados.

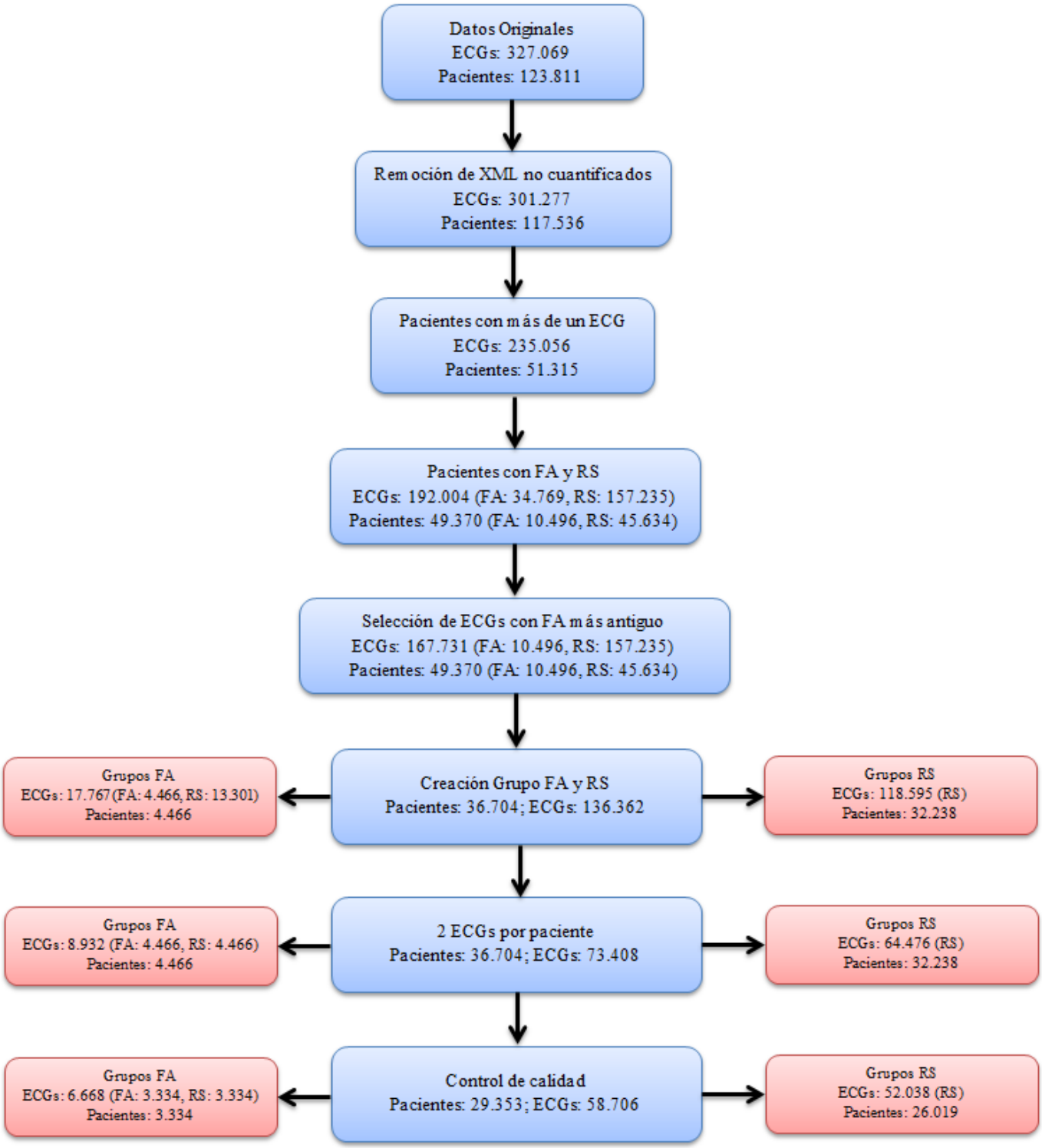

Figura 25. Diagrama del proceso de selección de pacientes y archivos XML a ser analizados.

Una vez obtenida la información filtrada, constituida por 2 dataframes correspondiente al grupo de pacientes que poseen un registro en FA con un registro en RS previo (grupo FA), y otro grupo de pacientes que poseen 2 registros en RS sin haber reportado en la base de datos FA (grupo RS). Se procedió a analizar dicha información. Puesto que la intención del proyecto es identificar variables y desarrollar modelos que permitan pronosticar la posibilidad de padecer FA a partir de ECG considerados normales, se compararon los valores cuantificados del trazado de las ondas del registro RS del grupo FA y el primer registro RS del grupo RS, considerando un rango de distancia similar entre registros del mismo grupo. Esta distancia se estableció con una media de 360 días y se puede extender a 360 días por debajo o por encima de esta media, arrojando un rango completo de 0 a 720 días (Figura 26).

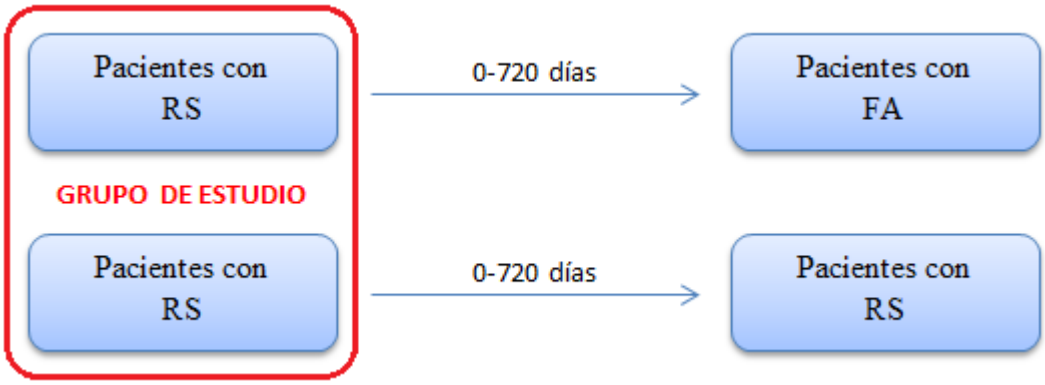

Figura 26. Diseño experimental para el estudio de los archivos XML (ECGs) del Hospital Universitario de La Princesa.

El análisis de los datos se realizó con dos acercamientos diferentes: El primero de ellos fue un análisis global donde se toman en consideración toda la muestra poblacional. Y el segundo acercamiento consistió en la segmentación de la población por grupos de edad y sexo. El objetivo de este segundo análisis consiste en determinar si:

- Segmentar la población permite disminuir el ruido en los datos e incrementar la capacidad predictiva de los modelos.
- Existe algún efecto en base a los grupos de edades.
- El sexo del paciente tiene algún efecto en el modelo y si existen indicadores que se funcionen mejor en base al sexo.

Los subconjuntos de edades empleados fueron los siguientes: 40 a 49, 50 a 59, 60 a 69, 70 a 79 y más de 80 años. En análisis se realizó para el conjunto completo de la población de cada grupo y para hombres y mujeres por separado. La base de datos se dividió en un 70% para el entrenamiento de los modelos y un 30% para la evaluación de los mismos.

Para cada grupo de edad y sexo se realizó una descripción y caracterización de la población y se garantizó que no existieran diferencias significativas entre los grupos FA y RS. Para las características de edad y distancia entre ECGs (días), se realizó la prueba t-test y para el sexo, en el caso del análisis del grupo completo (hombres y mujeres), se hizo la prueba de proporciones iguales. Luego se llevó a cabo un análisis univariante para determinar las variables que presentan diferencias significativas entre los grupos FA y RS empleando para ello la prueba Wilcoxon. Seguidamente se obtuvieron los p-values ajustados por el método de Bonferroni y se seleccionaron las variables con un p-value ajustado inferior a 0.05.

Con un análisis de covarianza se determinaron qué variables se encuentran correlacionadas mediante el método Spearman y se descartaron ya que pueden afectar negativamente al modelo. Por otra parte, ya que el tamaño del grupo RS es mucho mayor que el grupo FA (datos desbalanceados), fue necesario hacer un *"downsample"*, en el que se redujo aleatoriamente el tamaño del grupo RS. Esto fue necesario ya que al haber una desproporción en el tamaño de los grupos en comparación (FA y RS), los modelos que se generaran tenderían a predecir "preferentemente" al grupo de mayor tamaño dando lugar a modelos sesgados. Posteriormente, se emplearon los datos para la construcción de los modelos predictivos, optimizando los híperparámetros correspondientes a los algoritmos utilizados. Cada modelo se entrenó y evaluó 10 veces siendo los resultados finales la media de estas 10 mediciones. En cada una de las iteraciones se hizo un *"downsample"* diferente.

Uno de los métodos empleados para entrenar los modelos fue SVM ("*Support Vector Machine*"), empleando el paquete e1071 [25]. De este paquete se utilizaron las funciones:

- "svm": Función para entrenar el modelo.
- "tune": Función que permite hacer la optimización de los híperparámetros del modelo.

El otro método empleado para la construcción de modelos fue "*Extreme Gradient Boosting*" (XGBoost), del paquete "xgboost [26] junto con el paquete "caret" [27], del cual se usaron las siguientes funciones:

- "confusionMatrix": Calcula una tabulación cruzada de clases observadas y predichas con estadísticas asociadas.
- "createDataPartition": Para crear las particiones de los datos de prueba y entrenamiento.
- "downSample": Muestrea aleatoriamente un conjunto de datos para que todas las clases tengan la misma frecuencia que la clase minoritaria.
- "train": Esta función configura un conjunto de parámetros de ajuste para varias rutinas de clasificación y regresión, ajusta cada modelo y calcula un re-muestreo basada en medidas de rendimientos.

- “trainControl”: Controla los matices computacionales de la función “train”.

Alternativamente al análisis univariante con p-values ajustado y el análisis de covarianza se empleó el método “*Sequential Forward Floating Selection*” (SFFS) o en español Selección secuencial flotante hacia adelante para la selección de un conjunto de variables óptimo. Los algoritmos de selección de características secuenciales son una familia de algoritmos de “*greedy search*” que se utilizan para reducir un espacio de características d-dimensional iniciales a un subespacio de características k-dimensional donde $k < d$. La motivación detrás de los algoritmos de selección de características es escoger automáticamente un subconjunto de características que sean más relevantes para el problema. El objetivo principal de la selección de características es mejorar la eficiencia computacional y reducir el error de generalización del modelo eliminando características irrelevantes o ruido [28].

Las variantes flotantes o “*floating*”, pueden considerarse como extensiones de los algoritmos de selección de característica secuenciales simples. Los algoritmos flotantes tienen un paso adicional de exclusión o inclusión de características una vez que se incluyeron (o excluyeron), de modo que se pueda muestrear un mayor número de combinaciones de subconjuntos de características. Es importante enfatizar que este paso es condicional y solo ocurre si el subconjunto de características resultante es evaluado como "mejor" por la función de criterio después de la eliminación (o adición) de una característica particular [29]. Este método se empleó para el conjunto completo de datos, es decir para todas las edades y ambos sexos, usando los algoritmos de SVM y XGBoost.

Otros paquetes empleados:

- “dplyr”: Proporciona un conjunto de herramientas para manipular eficientemente conjuntos de datos en R [30].

- “lubridate”: Funciones para trabajar con fechas y horas. Permite el análisis, extracción, actualización y manipulación algebraica de datos de fecha y hora [31].

- “corrplot”: Para la visualización de la matriz de correlación [32].

- “psych”: De este paquete se usó la función corr.test que calcula las correlaciones, tamaños de muestra y valores de probabilidad entre elementos de una matriz o marco de datos [33].

4. RESULTADOS Y DISCUSIÓN

4.1. SEGMENTACIÓN DE LA POBLACIÓN POR GRUPO DE EDAD Y SEXO.

En la Tabla 2 se puede apreciar la constitución final de la población estudiada por grupos de edad y sexo. Se puede notar de forma general el incremento del número de casos de FA con el aumento de la edad (Figura 27). También se puede notar como el número de casos suele ser mayor en hombres que en mujeres, con excepción del grupo de pacientes mayores de 80 años, donde se hay un incremento más pronunciado en la población femenina.

Tabla 2. Tamaño de la población por grupos de edad y sexo.

		Subconjunto de Edades					
	Grupo	**40-49**	**50-59**	**60-69**	**70-79**	**>80**	**Todos**
Hombres	**FA**	36	93	168	304	421	966
	SR	828	1200	1469	1435	1337	6722
	SR (Ajustado)	613	888	1087	1062	989	3361
	Total	**649**	**981**	**2510**	**1366**	**1410**	**4327**
Mujeres	**FA**	23	36	91	242	548	906
	SR	726	1091	1137	1464	1955	6686
	SR (Ajustado)	537	807	841	1083	1447	3477
	Total	**560**	**843**	**932**	**1325**	**1995**	**4383**
Hombres y Mujeres	**FA**	59	129	259	546	969	1872
	SR	1554	2291	2606	2899	3292	13408
	SR (Ajustado)	1126	1356	1889	1931	2387	6167
	Total	**1185**	**1485**	**2148**	**2477**	**3356**	**8039**

Una vez segmentada la población por grupo de edad y sexo se procedió a verificar que no hubiera diferencias significativas entre los grupo FA y SR. En todos los casos se mostró que existían diferencias significativas entre los grupos. Ya que el grupo RS es mucho mayor que el grupo FA, se ajustó la población del grupo RS para disminuir las diferencias con el grupo FA con respecto la edad, sexo y la distancia entre ECGs. En la tabla 2 se puede observar la reducción del grupo RS después del ajuste. En la Figuras 28 y 29 se puede apreciar las gráficas de boxplot para los distintos grupos luego del ajuste con respecto la edad y la distancia entre ECGs respectivamente. Y en la Tabla 3, se encuentran los resultados del estudio estadístico en donde se muestra que ya no existen diferencias significativas entre los grupos FA y RS.

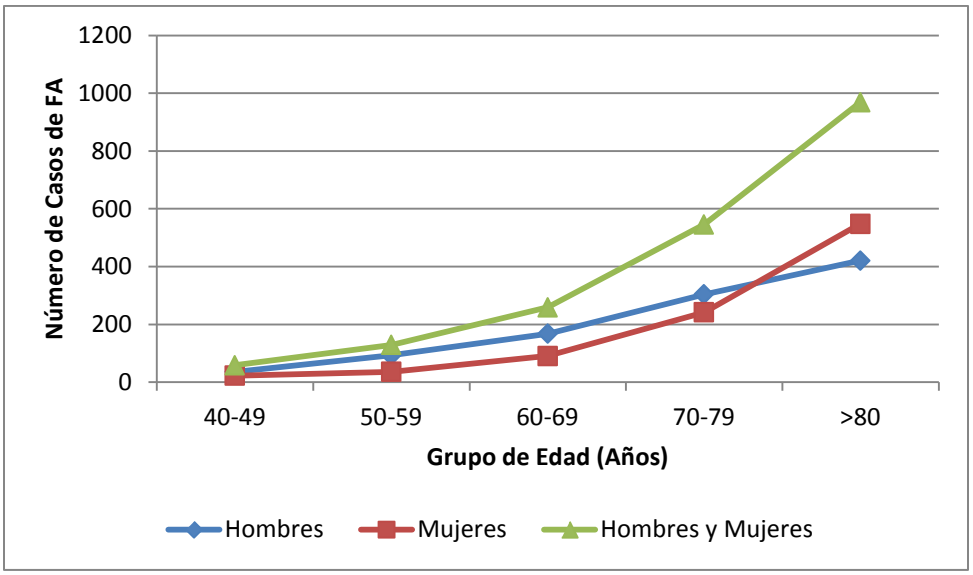

Figura 27. Número de casos de FA por grupo de edad y sexo en el análisis final.

Tabla 3. Valores estadísticos de la comparación de los grupo FA vs RS para los diferentes subconjunto de edad y sexo luego del ajuste.

	Medida		Subconjunto de Edades					
			40-49	**50-59**	**60-69**	**70-79**	**>80**	**Todos**
Hombres	Media edad (años)	FA	45.72	55.41	65.33	75.33	85.43	74.14
		RS	45.25	55.08	65.47	75.18	85.39	74.05
	p-value (t-test)		**0.1466**	**0.1859**	**0.4388**	**0.3172**	**0.8603**	**0.7967**
	Media distancia (días)	FA	311.64	320.87	351.58	352.83	352.57	332.07
		RS	301.15	319.75	327.93	342.68	342.20	328.58
	p-value (t-test)		**0.7794**	**0.9642**	**0.2104**	**0.4912**	**0.4376**	**0.678**
Mujeres	Media edad (años)	FA	45.22	55.06	65.34	75.08	86.83	79.24
		RS	45.45	55.01	65.47	74.86	86.48	78.84
	p-value (t-test)		**0.6385**	**0.9158**	**0.5894**	**0.163**	**0.05197**	**0.2176**
	Media distancia (días)	FA	399.30	334.33	359.87	361.60	339.18	341.30
		RS	338.17	337.67	364.99	349.65	330.30	336.44
	p-value (t-test)		**0.1854**	**0.9261**	**0.8277**	**0.4281**	**0.4197**	**0.5523**
Hombres y Mujeres	Media edad (años)	FA	45.53	55.31	65.33	75.22	86.22	76.61
		RS	45.36	55.03	65.46	75.38	86.06	76.26
	p-value (t-test)		**0.5866**	**0.1775**	**0.3776**	**0.1634**	**0.2242**	**0.1638**
	Media distancia (días)	FA	345.81	324.63	354.49	356.72	344.99	336.53
		RS	322.58	333.44	344.02	341.98	331.39	335.03
	p-value (t-test)		**0.4345**	**0.6696**	**0.479**	**0.1685**	**0.1106**	**0.801**
	Proporción (sexo)	FA	0.0428	0.0736	0.1038	0.2174	0.2802	0.2276
		RS	0.0556	0.0934	0.1322	0.2229	0.3007	0.2380
	p-value (equality of proportions)		**0.378**	**0.2411**	**0.0548**	**0.7823**	**0.2088**	**0.2785**

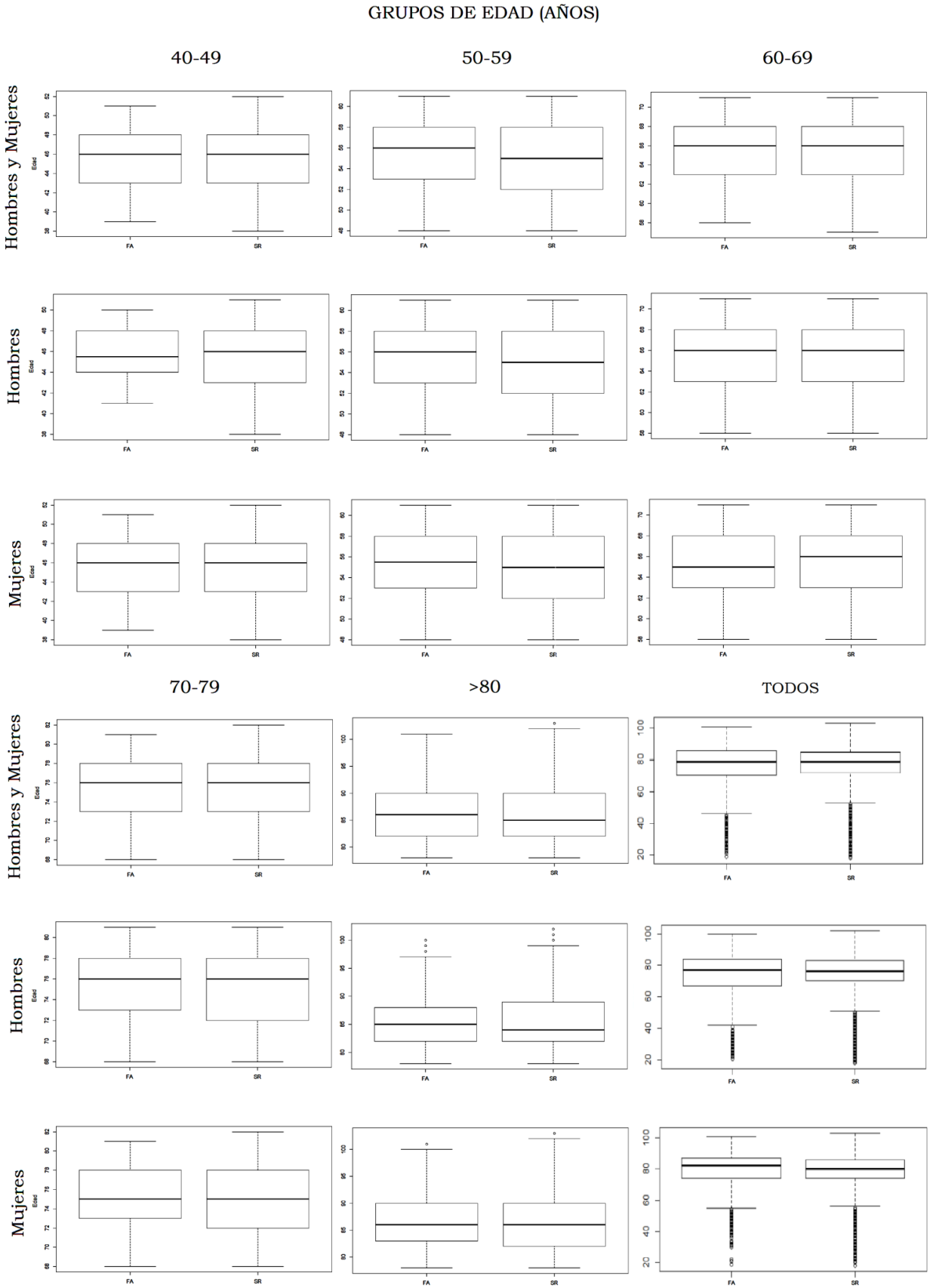

Figura 28. Gráficas Boxplot de los grupos FA y RS con respecto a la edad para los diferentes subconjuntos de edades luego del ajuste.

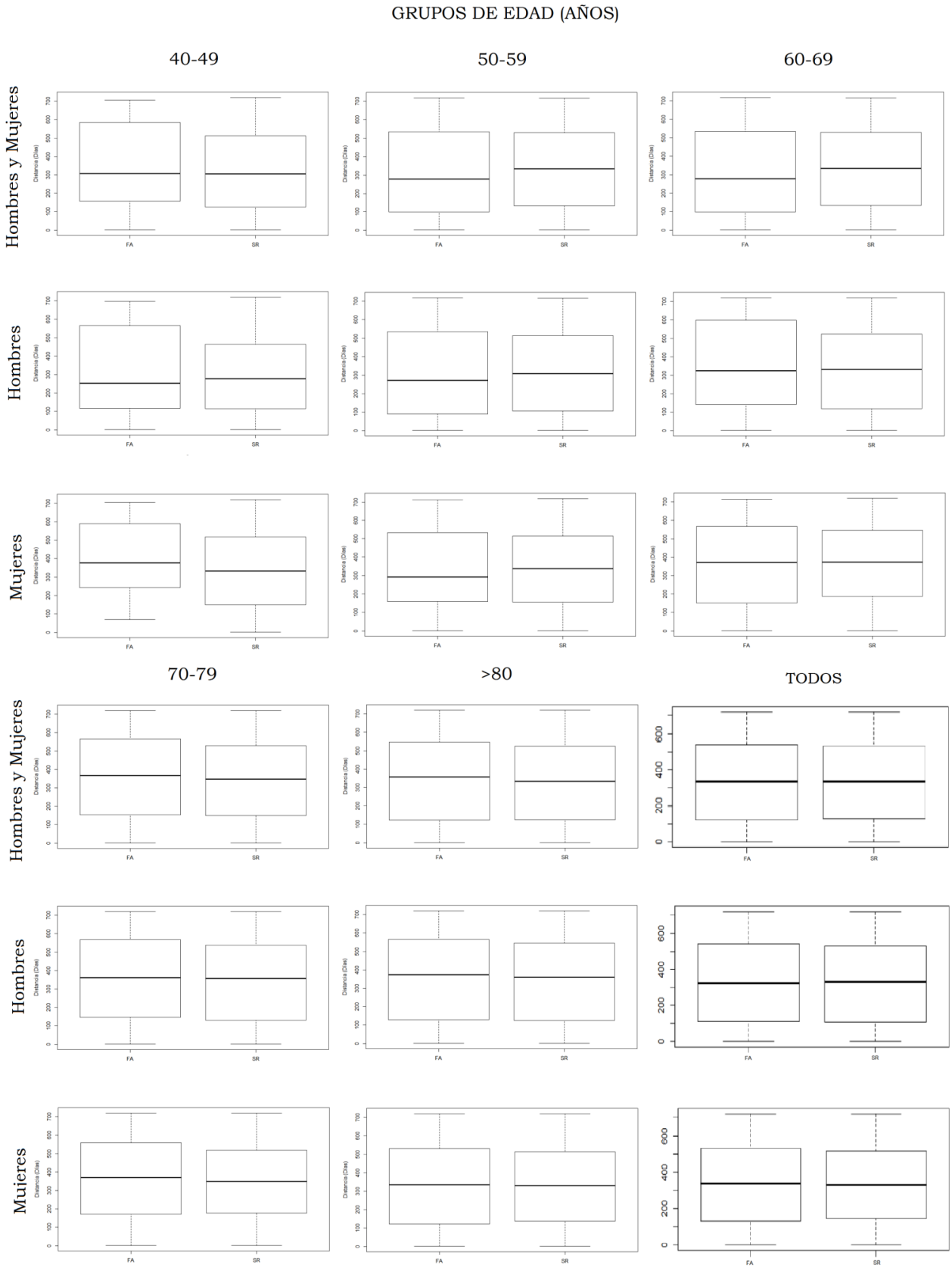

Figura 29. Gráficas Boxplot de los grupos FA y RS con respecto a la distancia entre ECGs para los diferentes subconjuntos de edades luego del ajuste.

4.2. DETERMINACIÓN DE VARIABLES SIGNIFICATIVAS Y CONSTRUCCIÓN DE MODELOS PREDICTIVOS.

Una vez establecido los subconjuntos de edades en donde los grupos FA y RS poseen características de edad, sexo y distancia entre ECGs similares. El siguiente paso fue realizar el análisis univariante determinado las variables significativas para cada subconjunto de edades y mediante un estudio de correlación se seleccionaron las variables a emplear para la construcción de los modelos. Dado que la prevalencia de la FA incrementa con la edad, el tamaño de la muestra se incrementará con la misma.

En el caso de los grupos de 40-59 años se encontraron pocas o ninguna variable con diferencias significativas como se puede observar en la Tabla 4, siendo el grupo más joven quizás tenga menos alteraciones en el ECG. Es importante destacar que parece existir una tendencia en el incremento de variables significativas con la edad. Este patrón cambia para el grupo de más 80 años, donde además se observa un incremento más marcado en los casos FA en mujeres (Figura 27). Por otra parte, el número de variables que presentan diferencias significativas es mayor en el grupo donde se incluyen todas las edades. Por ende, los resultados obtenidos parecen indicar que los marcadores varían según la edad y el sexo del paciente, aunque esta variación se muestra de forma más pronunciada para los grupos de edad que para el sexo. Además emplear grupos poblacionales más heterogéneos se traduce en un mayor número de variables y modelos más complejos.

En la tabla 4 también se puede notar que donde se suelen medir con mayor frecuencia las diferencias entre el grupo FA y RS son en las medidas grupales y en las derivaciones I y II, dentro de las cuales las variables que suelen ser más afectadas son la amplitud y duración de la onda P principalmente, pero también las ondas R, S, T y Q en menor medida. Indicando que posiblemente ya existiría un diferencia notable en la función de la aurículas y su efecto sobre los ventrículos antes de manifestarse la patología. También se observan variaciones en los intervalos PR, ST. En cuanto a las medidas grupales están presentes en casi todos los grupos edad y se pueden destacar las siguientes variables:

- “memberpercent”: Porcentaje del número total de latidos.
- “membercount”: Número de latidos.
- “atrialratestddev”: Desviación estándar de la frecuencia auricular.
- “printstddev”: Desviación estándar del intervalo PR.
- “avgpcount”: Número promedio de ondas P por QRS complejo.
- “highprint”: El intervalo PR más largo.
- “meanprint”: El intervalo PR medio.

Estas variables indican la presencia de alteraciones en el ritmo y la relación entre la onda P y la onda R. A partir de estos resultados se puede intuir que en ECGs considerados normales ya existirían indicativos de podrían pronosticar si un paciente podría padecer FA.

Tabla 4. Variables que muestran diferencias significativas entre el grupo FA y RS en cada una de los subconjuntos de edades.

	Variables	Subconjuntos de Edades					
		40-49	**50-59**	**60-69**	**70-79**	**>80**	**Todos**
Hombres	Significativas	3	6	10	39	29	41
	No correlacionadas	2	5	6	10	5	12
	Variables del modelo	aVR_tptpdur, Group_meanqtint	V3_rdur, V4_qrsarea, V6_sdur, Group_ventratestddev, Group_meanprint	aVR_pamp, V3_parea, Group_memberpercent, Group_meanqrsdur, Group_highprint, Group_printstddev	I_pamp, II_pamp, II_pdur, aVR_print, aVF_pdur, V1_pamp, Group_membercount, Group_memberpercent, Group_atrialratestddev, Group_printstddev	I_pamp, II_pamp, II_tamp, Group_atrialratestddev, Group_avgpcount	I_pamp, I_print, II_pamp, aVR_tamp, V1_ppppdur, V2_stslope, V3_parea, Group_membercount, Group_memberpercent, Group_atrialratestddev, Group_avgpcount, Group_printstddev
Mujeres	Significativas	N/A	N/A	10	51	22	61
	No correlacionadas	N/A	N/A	3	9	7	10
	Variables del modelo	N/A	N/A	I_pamp, I_ston, III_tptpdur	I_pamp, I_rdur, I_ston, II_pamp, II_print, aVL_rdur, V3_pdur, Group_lowventrate, Group_atrialratestddev	I_pamp, II_pamp, V1_pamp, Group_memberpercent, Group_atrialratestddev, Group_highprint, Group_printstddev	I_pamp, I_st80, II_pamp, II_ston, aVR_print, V1_pamp, V3_qrsppk, Group_memberpercent, Group_ventratestddev, Group_printstddev
Hombres y Mujeres	Significativas	4	3	21	77	36	98
	No correlacionadas	3	2	7	13	11	14
	Variables del modelo	aVR_tdur, Group_ventratestddev, Group_highprint	V6_sdur, Group_meanprint	I_pamp, I_qrsdur, II_pamp, aVF_print, V3_pppparea, V3_qrsarea, Group_memberpercent	I_pamp, I_ppppdur, I_rdur, I_tamp, I_print, II_pamp, II_st80, V1_pamp, V3_pamp, Group_membercount, Group_memberpercent, Group_atrialratestddev, Group_printstddev	I_pamp, II_pamp, aVR_ppppdur, V1_pamp, V4_sdur, Group_memberpercent, Group_ventratestddev, Group_atrialratestddev, Group_avgpcount, Group_highprint, Group_printstddev	I_pamp, I_rdur, I_stmid, I_print, II_pamp, II_ston, V1_pamp, V3_samp, V4_sdur, Group_membercount, Group_memberpercent, Group_atrialratestddev, Group_avgpcount, Group_printstddev

En la tabla 5 se pueden observar los resultados de la evaluación de los modelos predictivos con respecto a los subconjuntos de edades y sexo sobre los datos de prueba y entrenamiento. Estos resultados están representados empleando la medida de exactitud (*"Accuracy"*), que viene dada por la siguiente ecuación:

$$Exactitud = \frac{número\ de\ aciertos}{número\ de\ predicciones}$$

Tabla 5. Resultados de la evaluación de los modelos con respecto a los datos de entrenamiento y prueba representados en exactitud *("Accuracy")*.

			Subconjunto de Edades (Exactitud o *"Accuracy"*)					
		Grupo	**40-49**	**50-59**	**60-69**	**70-79**	**>80**	**Todos**
Hombres	XGBoost	Test	0.645	0.606	0.524	0.587	0.594	0.562
		Training	0.719	0.739	0.776	0.728	0.689	0.699
	SVM	Test	0.625	0.587	0.529	0.605	0.590	0.582
		Training	0.675	0.742	0.758	0.702	0.637	0.632
Mujeres	XGBoost	Test	N/A	N/A	0.591	0.576	0.588	0.582
		Training	N/A	N/A	0.787	0.776	0.699	0.701
	SVM	Test	N/A	N/A	0.604	0.585	0.592	0.599
		Training	N/A	N/A	0.731	0.663	0.650	0.625
Hombres y Mujeres	XGBoost	Test	0.585	0.536	0.575	0.582	0.606	0.590
		Training	0.775	0.676	0.721	0.747	0.722	0.682
	SVM	Test	0.553	0.509	0.604	0.596	0.591	0.590
		Training	0.848	0.664	0.666	0.658	0.675	0.632

Los valores de exactitud para los datos de entrenamiento se encuentran entre 0,62 y 0,85, mientras que para los datos de prueba están entre 0,50 y 0,65. Como es de esperar los resultados en los datos de entrenamiento presentan un mejor desempeño que en los datos de prueba, los cuales son datos "nuevos" o no conocidos para los modelos.

Con respecto a los algoritmos empleados para la construcción de los modelos se puede destacar que tanto en SVM como XGBoost presentaron resultados similares, alternándose de manera equitativa el mejor desempeño en los diferentes subconjuntos de datos, aunque las diferencias en las exactitudes entre los algoritmo para el mismo subconjunto de datos suelen ser pequeñas, entre 1-3% aproximadamente.

Comprando los resultados entre los diferentes grupos se puede observar que para el subconjunto de 70 a 79, más de 80 y todos, las exactitudes entre los grupos son muy

similares, al igual que al comparar entre los modelos para hombre, mujeres y ambos sexos en estos subconjunto, existiendo diferencias entre el 1-3%. Por el contrario para los subconjuntos de 40 a 49, 50 a 59 y 60 a 69 años se observa mayor variación en los resultados de exactitud tanto con respecto a la edad como el sexo. Esta mayor variabilidad puede ser debida a que las poblaciones en estos subconjuntos son menores que en el resto.

Con el propósito de mejorar los resultados obtenidos, se empleó el método SFFS para la selección de un subconjunto de variables óptimo alternativo. Dado que los resultados logrados previamente para el conjunto completo de datos y ambos sexos, de manera general, fueron similares a los resultados por subconjuntos de edades, el método SFFS solo se utilizó para la totalidad de la población. En la Figura 30 y 31 se pueden observar los resultados conseguidos, a partir de los cuales se seleccionó el mejor modelo en base a los resultados sobre los datos de validación.

El mejor modelo arrojó una exactitud del 75.4 %, 66.5 %, 60.7% para SVM y 72.3 %, 66.7 %, 60.9 % para XGBoost esto sobre conjunto de datos de entrenamiento, validación y prueba respectivamente. Dichos modelos emplearon 62 variables para SVM y 39 para XGBoost (Tabla 6). Cabe notar que a pesar del incremento considerable en el número de variables la mejoría en la exactitud no mostró muchas diferencias, obteniendo un incremento inferior al 2%.

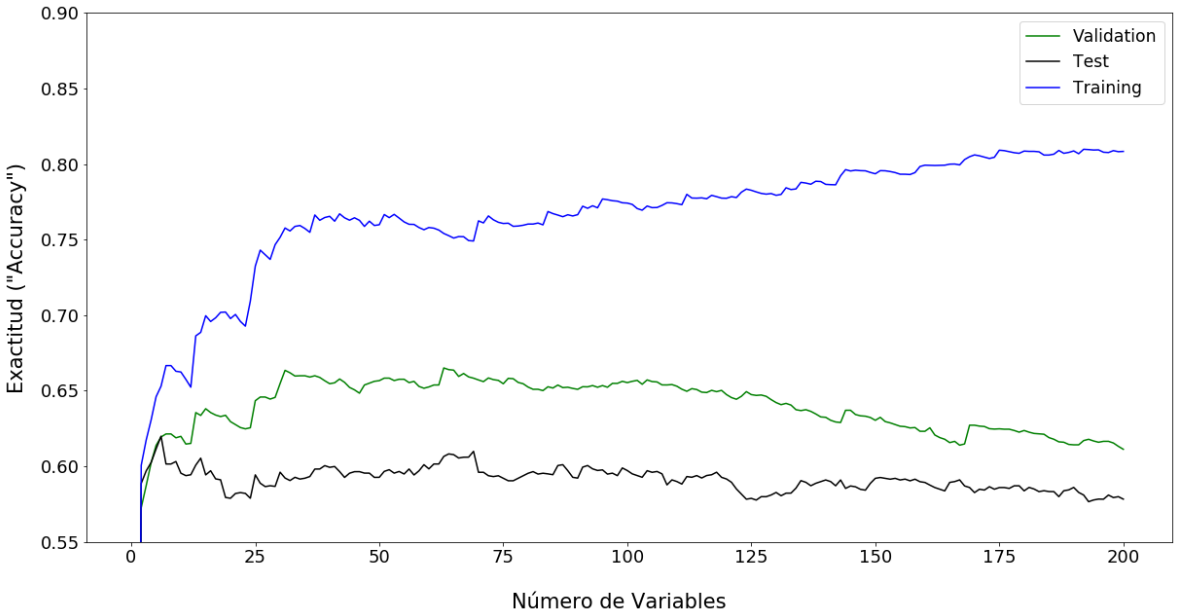

Figura 30. Representación gráfica de los resultados del estudio SFFS para la determinación del subconjunto de variables óptimo para el total de la población empleado el algoritmo SVM.

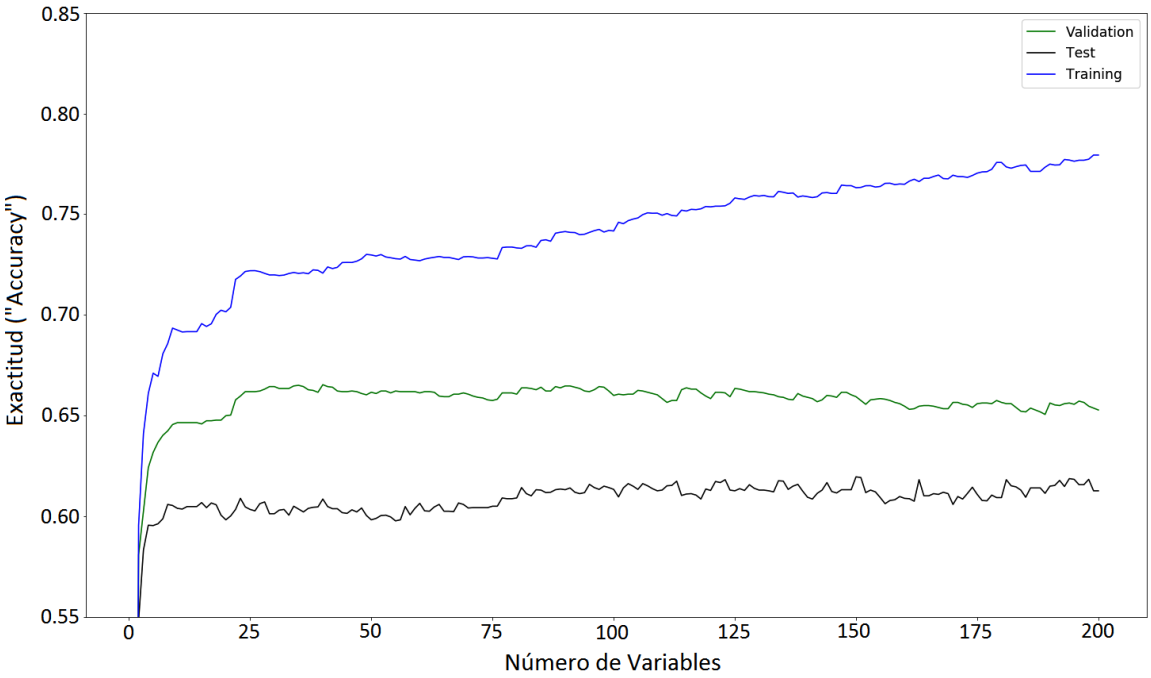

Figura 31. Representación gráfica de los resultados del estudio SFFS para la determinación del subconjunto de variables óptimo para el total de la población empleado el algoritmo XGBoost.

Tabla 6. Conjunto óptimo de variables determinadas por el método SFFS para la población total empleando los algoritmos SVM y XGBoost.

			Subconjunto de Edades
			Todos
Hombres y Mujeres	SVM	N° de Variables	62
		Variables del Modelo	I_pamp, I_rpamp, I_rpdur, I_spamp, I_spdur, I_stend, II_parea, II_pppparea, II_qamp, II_ramp, II_qrsarea, II_st80, 'III_samp, III_qrsarea, III_tparea, aVR_pamp, aVR_pppparea, aVR_spamp, aVR_st80, aVR_tpdur, aVL_ppdur, aVL_ppppdur, aVL_pparea, aVL_spamp, aVL_spdur, aVL_tparea, aVF_rpamp, aVF_qrsarea, V1_st80, V1_tamp, V1_tarea, V1_tpamp, V1_tpdur, V1_tptparea, V2_ppamp, V2_tpamp, V2_tparea, V3_pamp, V3_vat, V3_tparea, V4_parea, V4_pppparea, V4_qamp, V4_rpamp, V5_rpamp, V5_rpdur, V5_qrsarea, V5_tptparea, V6_pamp, V6_rpamp, V6_rpdur, V6_spamp, V6_spdur, V6_tpdur, Group_lowventrate, Group_meanventrate, Group_highventrate, Group_meanrrint, Group_atrialrate, Group_avgpcount, Group_meanqtint, Group_comppausecount
	XGBoost	N° de Variables	39
		Variables del Modelo	I_pamp, I_sdur, I_spamp, I_spdur, I_stslope, II_ppppdur, II_qdur, II_rpamp, II_spamp, II_spdur, II_tparea, III_pparea, III_tarea, aVR_pparea, aVR_qdur, aVR_spamp, aVR_print, aVF_spamp, aVF_spdur, V1_stslope, V2_spamp, V2_spdur, V2_tdur, V3_rpamp, V3_spamp, V3_spdur, V4_tpamp, V5_ppamp, V5_rpamp, V5_spamp, V5_spdur, V5_vat, V5_tpamp, V6_samp, V6_spamp, V6_spdur, Group_meanqrsdur, Group_meanprseg, Group_comppausecount

En general, los valores de exactitud obtenidos en este estudio se encuentran alrededor del 0,60 o 60%. Un aspecto importante a tener en consideración a la hora de evaluar estos resultados, es que la duración de los ECGs empleados en este trabajo son de tan solo 10 segundos. En estudios previos [34, 35, 36] se han obtenidos resultados por encima del 80%, evaluando ECGs con una duración desde minutos hasta varias horas. Estos ECGs fueron segmentados en diferentes tamaños (donde además se evaluó el efecto del tamaño del segmento), proporcionando mucha mayor información por paciente, lo que claramente puede a ver tenido un efecto en los resultados finales y la generación de modelos con un mayor poder de clasificación. Cabe destacar que estos estudios previos se realizaron con un tamaño de población mucho menor entre 50 y 150 pacientes.

5. CONCLUSIONES

CONCLUSIONES

- Se evaluó la estructura de los archivos XML, donde se encuentra almacenada la información de los ECGs, identificando la información de interés. Esta información incluye tanto la empleada para el análisis como la información sensible que pudiera identificar a los pacientes.

- Se realizó una anonimización de los archivos XML, mediante el desarrollo de un script en Bash.

- Se extrajo toda la información de interés (más de 444 variables) de cada uno de los más de 320.000 archivos XML de la base de datos anonimizada.

- La información fue depurada para eliminar errores en la base de datos y para seleccionar el grupo de estudio conformado por pacientes que solo presenta ECGs normales o en RS (grupo control) y pacientes en FA con RS previo (grupo de casos).

- Se realizó el segmentado de la población en grupo de edad y sexo, a partir de los cuales se construyeron los modelos predictivos empleando los algoritmos SVM y XGBoost, mostrando resultados similares.

- Los resultados para el subconjunto de edades de 70 a 79, más de 80 y total, no parecen mostrar una diferencia notable en la exactitud debido a la edad o el sexo. A diferencia de los subconjuntos de 40 a 69 años, donde sí se observó una mayor variabilidad.

- Se determinaron algunas posibles características o variables discriminantes entre FA y RS para los distintos grupos de edad y sexo.

- Se empleó el método SFFS para la selección de un nuevo conjunto de variables y la construcción de modelos predictivos con SVM y XGBoost, obteniendo una mejoría inferior al 2% a costa de modelos más complejos.

- En general se obtuvieron modelos con una exactitud que ronda el 60%.

CONCLUSIONS

- The structure of the XML files was evaluated, which is where the information of the ECGs is stored, identifying the information of interest. This information includes both the one used for the analysis and the sensitive information that could identify the patients.

- Anonymization of the XML files was made, by developing a script in Bash.

- All the information of interest (more than 444 variables) was extracted from each of the more than 320,000 XML files in the anonymized database.

- The information was filtered to eliminate errors in the database and to select the study group consisting of patients who only have normal ECGs or SR (control group) and patients with AF with previous SR (case group).

- Segmentation of the population by age and sex was performed, from which the predictive models were constructed using the SVM and XGBoost algorithms, showing similar results.

- The results for the subset of ages 70 to 79, over 80 and total, do not seem to show a noticeable difference in accuracy due to age or sex. Unlike the subsets of 40 to 69 years, where a higher variability was observed.

- Some possible discriminant characteristics or variables between AF and SR were determined for the different age groups and sex.

- The SFFS method was used for the selection of a new set of variables and the construction of predictive models with SVM and XGBoost, obtaining an improvement of less than 2% at the expense of more complex models.

- Overall, models obtained had an accuracy of around 60%.

6. BIBLIOGRAFÍA

[1] Albert C. M. y Stevenson W. G. (2016). The Future of Arrhythmias and Electrophysiology. *Circulation, 133*(25), 2687–2696.

[2] Cardiac Arrhythmias (Febrero, 2019). Recuperado de https://www.health.harvard.edu/a_to_z/cardiac-arrhythmias-a-to-z

[3] Newman, T. (2018). The heart: All you need to know. Recuperado de https://www.medicalnewstoday.com/articles/320565.php

[4] Noble R.J., Hillis J.S., Rothbaum D.A. (1990) Electrocardiography. En: Walker H.K., Hall W.D. y Hurst J.W. (Edits.). *Clinical Methods: The History, Physical, and Laboratory Examinations*. (Tercera ed.). Boston: Butterworths.

[5] Ashley E.A. y Niebauer J. (2004). *Cardiology Explained*. London: Remedica; 2004. Chapter 3, Conquering the ECG.

[6] ECG Leads (s.f.) Recuperado de https://www.sciencedirect.com/topics/medicine-and-dentistry/ecg-leads

[7] Dang D., Arimie R., and Haywood L.J. (2002). A review of atrial fibrillation. *Journal of the National Medical Association. 94*(12), 1036-1048.

[8] Atrial fibrillation (s.f.). Recuperado de https://www.mayoclinic.org/diseases-conditions/atrial-fibrillation/symptoms-causes/syc-20350624

[9] Nesheiwat Z., Goyal A. y Jagtap M. (2019). Atrial Fibrillation (A Fib). Recuperado de https://www.ncbi.nlm.nih.gov/books/NBK526072/

[10] Atrial fibrillation (s.f.). Recuperado de https://stanfordhealthcare.org/medical-conditions/blood-heart-circulation/atrial-fibrillation/causes.html

[11] Atrial fibrillation (s.f.). Recuperado de https://www.nhs.uk/conditions/atrial-fibrillation/complications/

[12] Szeredi P., Lukácsy, G., Benkö T. & Nagy Z. (2014). *The semantic web explained the technology and mathematics behind web 3.0*. Nueva York: Cambridge University Press.

[13] Miller J. A. (2019). How Predictive Analytics Is Impacting Patient Care. Recuperado de https://healthtechmagazine.net/article/2019/10/how-predictive-analytics-impacting-patient-care-perfcon

[14] Patil A. (2019). AI and predictive analytics lead to improved delivery of healthcare services. Recuperado de https://www.healthcarebusinesstech.com/ai-and-predictive-analytics-lead-to-improved-delivery-of-healthcare-services/

[15] Cristianini N., y Shawe-Taylor, J. (2000). An Introduction to Support Vector Machines and Other Kernel-based Learning Methods. Cambridge: Cambridge University Press.

[16] Chen L. (2019). Support Vector Machine — Simply Explained. Recuperado de https://towardsdatascience.com/support-vector-machine-simply-explained-fee28eba5496

[17] Chen, T.Q. y Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. 785-794. arXiv:1603.02754v3.

[18] Dahan H., Cohen S., Rokach L. y Maimon O. (2014). Proactive Data Mining with Decision Trees. London: Springer.

[19] Njeri R. (2017). A Brilliant Explanation of Decision Tree Algorithms. Recuperado de http://www.acheronanalytics.com/acheron-blog/brilliant-explanation-of-a-decision-tree-algorithms

[20] Nishida K.(2017). Introduction to Extreme Gradient Boosting in Exploratory. Recuperado de https://blog.exploratory.io/introduction-to-extreme-gradient-boosting-in-exploratory-7bbec554ac7

[21] The Philips 12-Lead Algorithm Physician's Guide, Philips MedicalSystems. Publication M5000-91000, 1st ed. 2003.

[22] Lang D. T. (2019). XML: Tools for Parsing and Generating XML Within R and S-Plus. Versión del paquete de R: 3.98-1.20. https://CRAN.R-project.org/package=XML

[23] Gómez-Doblas J.J., Muñiz J., Martin J.J.A., Rodríguez-Roca G., Lobos J.M., Awamleh P., Permanyer-Miralda G., Chorro F.J., Anguita M., Roig E., en representación de los colaboradores del estudio OFRECE. (2014). Prevalencia de fibrilación auricular en España. Resultados del estudio OFRECE. *Rev Esp Cardiol, 67*(4), 259–269.

[24] What is Prevalence? (s.f.). Recuperado de https://www.nimh.nih.gov/health/statistics/what-is-prevalence.shtml

[25] Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F., Chang C. y Lin C. (2019). e1071: Misc Functions of the Department of Statistics, Probability Theory Group. Versión del paquete de R: 1.7-3. https://CRAN.R-project.org/package=e1071

[26] Chen T, He T., Benesty M., Khotilovich V., Tang Y., Cho H., Chen K., Mitchell R., Cano I., Zhou T., Li M., Xie J., Lin M., Geng Y. y Li Y. (2019) xgboost: Extreme Gradient Boosting. Versión del paquete de R: 0.90.0.2. https://CRAN.R-project.org/package=xgboost

[27] Kuhn M., Wing J., Weston S., Williams A., Keefer C., Engelhardt A., Cooper T., Mayer Z., Kenkel B., Benesty M., Lescarbeau R., Ziem A., Scrucca L., Tang Y., Candan C. y Hunt T. (2019). caret: Classification and Regression Training. Versión del paquete de R: 6.0-84. https://CRAN.R-project.org/package=caret

[28] Raschka, Sebastian (2018) MLxtend: Providing machine learning and data science utilities and extensions to Python's scientific computing stack. J Open Source Softw 3(24).

[29] Pudil P., Novovičová J., y Kittler J. (1994). Floating search methods in feature selection. *Pattern Recognition Letters, 15,* 1119–1125.

[30] Wickham H., François R., Henry L. y Müller K. (2019). dplyr: A Grammar of Data Manipulation. Versión del paquete de R: 0.8.3. https://CRAN.R-project.org/package=dplyr

[31] Spinu V., Grolemund G., Wickham H., Lyttle I., Constigan I., Law J., Mitarotonda D., Larmarange J., Boiser J. y Lee C. H. (2018). lubridate: Make Dealing with Dates a Little Easier . Versión del paquete de R: 1.7.4. https://CRAN.R-project.org/package=lubridate

[32] Wei T., Simko V., Levy M., Xie Y., Jin Y. y Zemla J. (2017). corrplot: Visualization of a Correlation Matrix. Versión del paquete de R: 0.84. https://CRAN.R-project.org/package=corrplot

[33] Revelle W. (2019). psych: Procedures for Psychological, Psychometric, and Personality Research. Versión del paquete de R: 1.8.12. https://CRAN.R-project.org/package=psych

[34] Urtnasan E., Kim, H., Park J., Kang D. y Lee K. (2019). Automatic Prediction of Atrial Fibrillation Based on Convolutional Neural Network Using a Short-term Normal Electrocardiogram Signal. *Journal of Korean Medical Science, 34*(7).

[35] Nuryani N., Harjito B., Yahya I. y Lestari A. (2011). Atrial Fibrillation Detection Using Support Vector Machine. *AUTOMATIKA, 52*(1), 58-67.

[36] Sovilj S., Oosterom A., Rajsman G. y Magjarevic R. (2010). ECG-based prediction of atrial fibrillation development following coronary artery bypass grafting. *Physiological measurement, 31*(5), 663-677.